The genetic and ecological landscape of plasmids in the human gut

Michael K. Yu[1,‡,*], Emily C. Fogarty[2,3,‡], A. Murat Eren[2,4,5,6,7,*]

[1]Toyota Technological Institute at Chicago; Chicago, IL 60637, USA; [2]Department of Medicine, University of Chicago; Chicago, IL 60637, USA; [3]Graduate Program in the Biological Sciences, University of Chicago; Chicago, IL 60637, USA; [4]Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory; Woods Hole, MA 02543, USA; [5]Alfred Wegener Institute for Polar and Marine Research, 27570 Bremerhaven, Germany; [6]Institute for Chemistry and Biology of the Marine Environment, University of Oldenburg, 26129 Oldenburg, Germany; [7]Helmholtz Institute for Functional Marine Biodiversity, 26129 Oldenburg, Germany

[‡] These authors contributed equally to this work.

[*] Correspondence: mikeyu@ttic.edu, meren@uchicago.edu

Abstract

Despite their prevalence and impact on microbial lifestyles, ecological and evolutionary insights into naturally occurring plasmids are far from complete. Here we developed a machine learning model, PlasX, which identified 68,350 non-redundant plasmids across human gut metagenomes, and we organized them into 1,169 evolutionarily cohesive 'plasmid systems' using our sequence containment-aware network partitioning algorithm, MobMess. Similar to microbial taxa, individuals from the same country tend to cluster together based on their plasmid diversity. However, we found no correlation between plasmid diversity and bacterial taxonomy. Individual plasmids were often country-specific, yet most plasmid systems spanned across geographically distinct human populations, revealing cargo genes that likely respond to environmental selection. Our study introduces powerful tools to recognize and organize plasmids, uncovers their tremendous diversity and intricate ecological and evolutionary patterns in naturally occurring habitats, and demonstrates that plasmids represent a dimension of ecosystems that is not explained by microbial taxonomy alone.

Main

As a class of mobile genetic elements[1], plasmids can occur in cells from all domains of life[2,3], typically as extrachromosomal and circular DNA. Plasmids replicate semi-independently of their hosts, and often transfer between cells as a mechanism of horizontal gene transfer[4–8]. A hallmark of plasmids is their remarkably diverse capacity to impact their microbial hosts through fitness-determining functions they carry[4–6], such as antibiotic resistance genes[9,10] and virulence factors[11,12]. Plasmids also exhibit many interesting genetic properties, such as frequent recombination, which can result in plasmids sharing recurrent “backbone” sequences but differing in their cargo genes[13–15]. These backbone sequences often encode for core replication and transfer machinery[13,14,16–18] that determine the set of compatible hosts they can inhabit[18,19] as well as regulate their copy number in a specific host[20]. Experiments in model systems and organisms in culture have revealed the critical impact of plasmids in microbial phenotypes and survival especially for pathogens with medical significance. Yet, our understanding of the diversity, ecology, and genetic architecture of naturally occurring plasmids are far from complete.

Recent advances in metagenomics offer unprecedented access to the entire DNA content of an environment without the need for cultivation[21]. In particular, metagenomic assembly and binning strategies have enabled the reconstruction and characterization of microbial genomes de novo[22], including those in the human gut[23] where microbes have been associated with health and disease states[24,25]. Metagenomic approaches have also been applied to study plasmid content[26], but such applications have been limited to shotgun sequencing of plasmid-enriched samples[27–29] or to surveying only a handful of metagenomes at a time[30–32]. Over the past decade, the number of publicly available metagenomes has rapidly increased, creating an opportunity to conduct large-scale studies to characterize the diversity of naturally occurring plasmids in complex ecosystems.

Several computational strategies have been developed to identify plasmids in sequence collections[17,33,34]. Yet, the ability to distinguish plasmids from bacterial chromosomes or from other mobile genetic elements such as viruses via computational strategies remains a challenge[35]. Popular plasmid prediction strategies rely on k-mer patterns learned from reference plasmid sequences[30,31,36], exploit known functions such as replication or conjugation genes[37–39], or use a combination of these features[34]. While these features can help identify plasmids similar to those in public databases, they are of limited utility to recognize novel plasmids. Other approaches focus on circularity of sequences during (meta)genomic assembly[32,40,41]; however, this strategy

overlooks plasmids that are linear, integrated, or found as assembly fragments, and may confuse other types of circular mobile elements for plasmids.

Here, we present PlasX, a machine learning approach to identify plasmids in complex microbial ecosystems, and MobMess, a robust network partitioning algorithm to gain insights into plasmid evolution at scale. Using PlasX we identified a collection of 68,350 non-redundant plasmids in the human gut microbiome that were more genetically diverse than reference plasmids and substantially more prevalent across global human populations. We then used MobMess to organize predicted plasmids into 'plasmid systems' based on shared backbone sequences, which provided us with an evolutionary framework to investigate plasmid cargo gene content as a function of environmental pressures.

Results

A plasmid classification system based on *de novo* gene families

To train our machine learning model, we first compiled a reference set of 16,827 plasmids and 14,367 chromosomal sequences from public databases (Figure 1A, Table S1), in which we identified 51.2 million open reading frames. We were able to assign a function to 71% of plasmid genes using the Cluster of Orthologous Genes (COG)[42] and/or Pfam[43] databases. In parallel we clustered all genes into 1,090,132 *de novo* gene families, which accounted for 95% of all plasmid genes (Figures 1B and S1A) and constituted the primary data for the training of PlasX (Figures 1, S1B, and S1C; Supplementary Information).

PlasX is a logistic regression which assigns a positive or negative coefficient to gene families that are likely to originate from sequences that are of plasmid or non-plasmid origin (Figure 1C). The algorithm predicts whether a given sequence is a plasmid or not by considering the coefficients of all gene families in the sequence, and assigns a score to each prediction that ranges between 0 to 1, where a prediction score over 0.5 suggests that the sequence is more likely to be a plasmid than not. To benchmark PlasX we compared its performance to three state-of-the-art algorithms, PlasClass[31], PPR-Meta[44], and Platon[45] using a cross-validation strategy in which we trained models on a subset of non-redundant reference sequences and evaluated their performance on the remaining sequences that were not used for training (Supplementary Information). PlasX

95 achieved the highest area under the precision-recall curve (AUCPR=0.70) with a substantial improvement compared to the next best method (Platon, AUCPR=0.23) (Figure 1D). We then conducted a test using the 21,012 new plasmids that were recently included in a large plasmid database, PLSDB[46]. PlasX correctly identified 81.5% (17,128) of these new sequences as plasmids, while Platon, the next best method in our cross-validation tests, only predicted 37.4%
100 (7,860) of them as plasmids in its most sensitive mode (Table S2). We further evaluated the performance of PlasX and other methods on a recently characterized novel plasmid of Wolbachia, pWCP[47]. PlasX was able to predict pWCP as a plasmid (score = 0.73), while none of the other methods in our tests recognized pWCP as a plasmid (Table S3). Finally, as an additional benchmarking step to determine PlasX's ability to distinguish plasmids from other mobile genetic
105 elements, we ran PlasX on all ICE sequences from the ICEberg database[48] (n=552) and all prophage sequences from the NCBI viral database (n=445). PlasX correctly classified 92.2% of ICEs as not plasmids (Table S4), and 93.2% of NCBI viral database as not plasmids (Table S5). Platon could also distinguish prophages from plasmids with a high accuracy of 99.6%, but its classification accuracy was much lower compared to PlasX's for ICEs, as Platon classified 37.1%
110 of ICEs as plasmids.

The improved efficiency of PlasX comes from its reliance on de novo gene families rather than sequence features or gene functions alone. Indeed, genes in microbial sequence collections often cannot be annotated to a known function, and a known function often groups together genes with
115 large sequence differences. By partitioning genes into homologous groups, de novo gene families both increase the fraction of input data usable for training and increase the resolution of the resulting units that improve training and prediction. For instance, the Pfam PF10609 is a broad family of genes related to *parA*, a gene that drives the partitioning of not only chromosomes[49] but also plasmids[50] during cell division. As genes that resolve to this function are found on 35% of
120 plasmids and 95% of chromosomes, it has no ability to distinguish plasmids and chromosomes (coefficient of -0.023). However, PF10609 in our dataset could be subdivided into two de novo gene families – a plasmid specific one (coefficient +0.455), and a chromosome-specific one (coefficient -0.198). Indeed, the gene sequences in the two de novo gene families of PF10609 show a clear divergence of plasmids and chromosomes into monophyletic groups (Figures 1F,
125 1G). Furthermore, 35.5% (398,174) of *de novo* gene families did not resolve to a known function, even though many of them had highly positive coefficients. In fact, 12,076 of gene families with unknown functions had coefficients over 0.1, and of the 200 gene families with the highest coefficients, 129 did not have any functional annotation. Our survey of gene families (Table S6)

highlights the difficulty of using functional annotations alone to infer the importance of a gene
130 family in plasmid recognition. While some gene families included keywords such as “plasmid”, “replication”, or “conjugation” in their functional descriptions to offer naive confirmations for plasmid relevance, most others were difficult to immediately associate with plasmids. For instance, the gene family with the 17th highest PlasX coefficient of 1.678 in our list was a family of lipoproteins (PF05714). Its annotation, “*Borrelia burgdorferi virulent strain associated*
135 *lipoprotein*” does not explicitly associate it with plasmids. Yet, it occurred in 168 plasmids and only 2 chromosomes in our training data and has been studied previously for conferring virulence in plasmids[51,52], suggesting that PlasX-assigned coefficients offer an effective means to identify key gene families to recognize plasmids.

140 Overall, these results show that PlasX performs better than state-of-the-art plasmid prediction approaches in cross-validation tests, is able to recognize novel plasmids that are not present in existing databases, and is less likely to confuse other mobile genetic elements with plasmids.

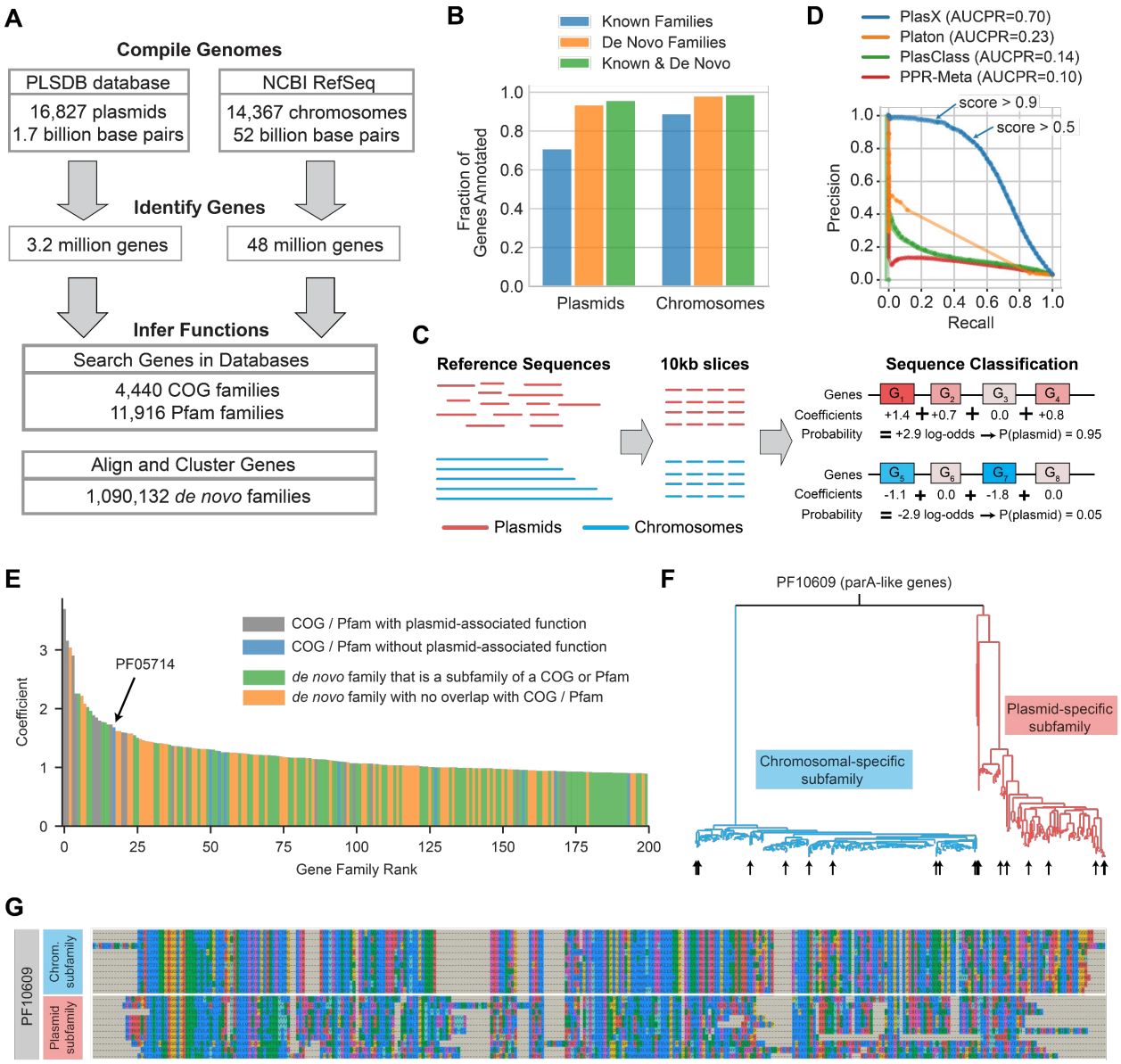


Figure 1. A machine learning model for classifying plasmids. (A) Our pangenomics workflow to characterize gene functions in a reference set of plasmids and chromosomes. **(B)** The fraction of all plasmids or all chromosomal genes that are annotated by using known families (blue), *de novo* families (orange), or a combination of both (green). **(C)** Training of PlasX. Reference sequences are sliced into 10 kbp windows and then prediction scores are made by a logistic regression that sums the contributions of gene families within a sequence. **(D)** Precision-recall curves comparing PlasX, Platon, PlasClass, and PPR-Meta. Except for PPR-Meta, every method was trained and evaluated using 4-fold cross-validation and an informed split. AUCPR was calculated using sequence weights for normalization. The arrows indicate the performance of PlasX using a score threshold of either >0.5 or >0.9. **(E)** The 200 gene families with the highest PlasX coefficients and thus most important for identifying plasmids. Gene families are ranked by their coefficient. **(F)** Maximum-likelihood phylogenetic tree of genes that are in PF10609 and also in either the plasmid-specific *de novo* subfamily mmseqs_5_1535552 (red) or chromosome-specific *de novo* subfamily mmseqs_70_40217271 (blue). **(G)** Sequence alignment of 10 representative genes from each subfamily (arrows in F).

PlasX unveils a large database of new plasmids from the human gut microbiome

Next, we applied PlasX to survey naturally occurring plasmids in the human gut microbiome, an environment which harbors a diverse range of microbes and mobile genetic elements[53]. For this, we assembled 36 million contigs from 1,782 human gut metagenomes, spanning culturally and geographically distinct human populations (Table S7). Running PlasX on these data resulted in a total of 226,194 predicted plasmids with a score above 0.5 (Figures 2A and S2A, Table S8). Our predictions spanned a wide range of lengths, including 135 sequences that were longer than 100 kbp, but they were generally shorter than reference plasmids with a median length of 2.6 kbp versus 53.3 kbp, respectively (Figure S2B). Part of this discrepancy is most likely due to the fragmented nature of assembled sequences from metagenomes. Indeed, the median length of the entire set of contigs was 2.1 kbp, and only 50,310 (0.14%) contigs were longer than 100 kbp. To minimize the impact of assembly fragments in our results, we removed predictions that did not appear to be circular and, at the same time, appeared to be fragments of more complete predictions in our collection. This filter left us with 100,719 predictions for downstream analyses (see Methods, Figure S3). Although this filtered set inevitably contains fragmented plasmids due to the nature of the input data, hereafter we refer to them as 'plasmids' for practical reasons.

To determine the circularity of plasmid sequences in metagenomes, we analyzed the orientation of mapped metagenomic paired-end reads (Figure 2B). With this approach we found that 19,652 plasmid sequences were circular, and we designated them as high-confidence plasmids for downstream analyses. Circular plasmids had a median lenth of 4.4 kbp and included sequences that were longer than 25 kbp (n=854), 50 kbp (n=378), or 100 kbp (n=47). An additional 14,151 sequences were not circular themselves but were highly similar to a circular sequence. Together, these two types of sequences defined a set of ‘circular-associated’ sequences representing 33.6%, or 33,803 of 100,719, of the predictions. Multiple factors can explain the lack of signal for circularity for the remaining plasmids, including insufficient sequencing depth to observe a sufficient number of reverse-forward pairs, fragmented contigs, or the non-circular nature of some plasmids that occur linearly[54] or those that are integrated into chromosomes[3].

Beyond circularity, confirming *in silico* whether a novel sequence represents a plasmid is a significant challenge. While single-copy core genes have been used to assess the completeness of non-plasmid and non-viral genomes assembled from metagenomes[22], our understanding of the

canonical features of plasmids is limited to a relatively small set of well-studied genes that are primarily derived from plasmids of model organisms in culture[37,38]. For instance, MOB-suite[38] identified canonical features for plasmid replication and conjugation in only 61% of the reference plasmid subtypes we used to train PlasX, which reveals the limits of conventional approaches to identify plasmid features and survey novel plasmids (Table S1). Indeed, MOB-suite identified canonical features in only 10.1% of our predictions (Table S8). Given this narrow sensitivity, we developed orthogonal data-driven strategies to increase confidence in our predictions (Supplementary Information). We found that 49.4% (49,739) of predictions had orthogonal support for being a plasmid, by MOB-suite or other metrics (Figure 2C), and 28.5% (28,658) had such support and were also novel (Table S8). Overall, these findings suggest that our collection of predicted plasmids include not only sequences that match known plasmids but also novel ones that can further advance our ability to infer the gene pool and ecology of naturally occurring plasmids.

While conducting experiments is the most reliable strategy for validation, the labor-intensive and low-throughput nature of such investigations represent a significant limit to their scale. Nevertheless, to experimentally validate at least some of our metagenome-derived predictions as true plasmids of the human gut, we developed a pipeline to identify predictions that are (1) present in human gut microbial isolates, (2) are circular in those isolates, and (3) can be naturally transferred to other microbes. First, we detected 127 of our predicted plasmids in 14 *Bacteroides* isolate genomes that we sequenced in a previous study[55] (Figure S4). For two of these plasmids, pFIJ0137_1 and pENG0187_1, we performed additional short-read and long-read sequencing to obtain complete plasmid genomes and confirm their circular configuration (Figure S5). Finally, we demonstrated the ability of pFIJ0137_1 to transfer as a plasmid and confer antibiotic resistance from one *B. fragilis* host to another (see Methods, Figure S4). While not comprehensive, these experimental results show that PlasX is able to predict novel plasmids that have canonical features of being extrachromosomal, circular, and transmissible between cells.

Overall, our survey of individual assemblies of human gut metagenomes using PlasX resulted in 100,719 plasmid sequences for in-depth characterization.

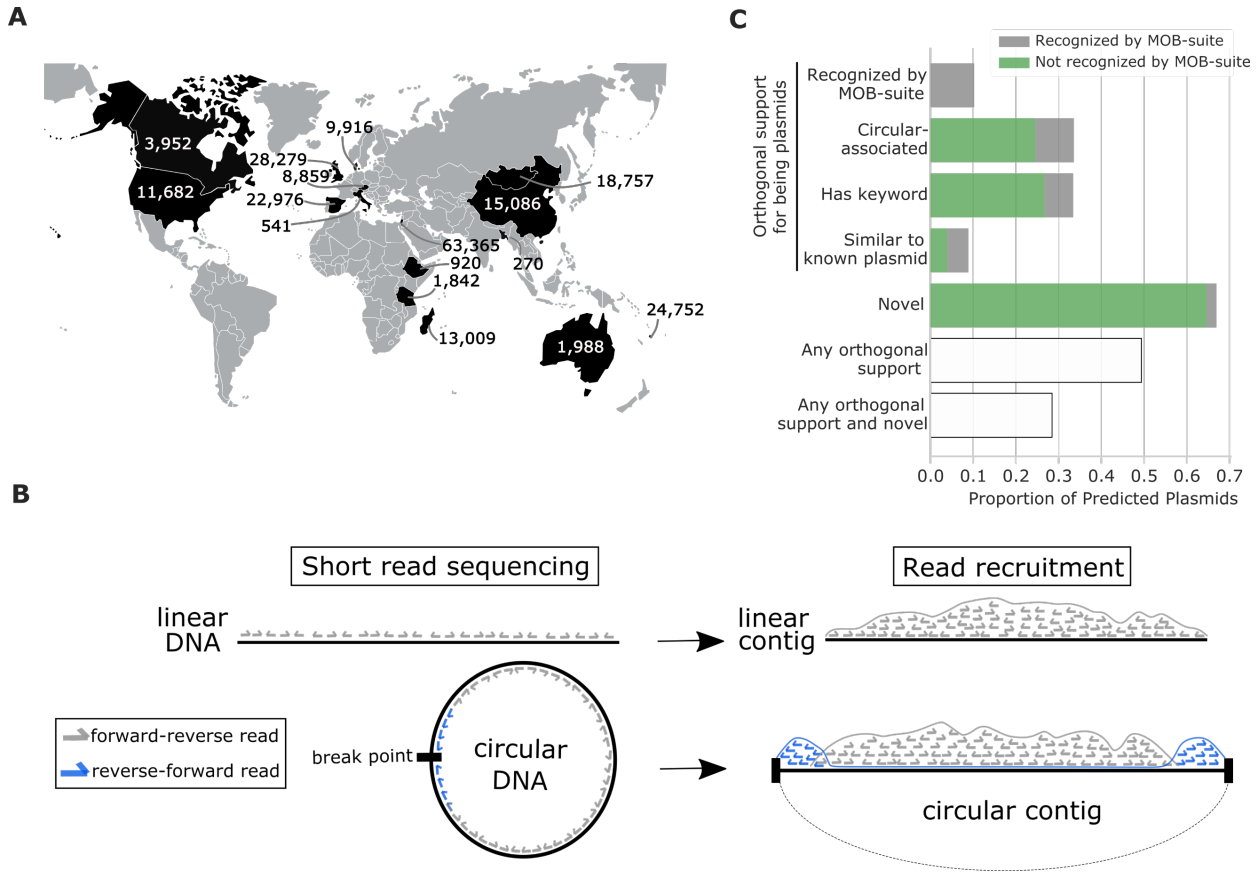


Figure 2. Plasmid prediction from metagenomes. (A) Number of plasmids predicted from each country. **(B)** Diagram of paired-end reads mapping to a linear versus a circular contig. Linear contigs have forward-reverse reads only, while circular contigs also have reverse-forward reads concentrated on the ends due to an artifact in contig assembly. **(C)** Orthogonal support for and novelty of the 100,719 non-fragment predictions. We performed several types of analyses to assess how many predictions are true plasmids or novel sequences. While most predictions were not recognized by MOB-suite (gray bars), we found that 24,689 of these (24.5% of the 100,719 non-fragment predictions) were circular-associated sequences. 26,921 (26.7%) were 'keyword-recognizable', as they contained a COG or Pfam function with the words 'plasmid' or 'conjugation'. And finally, 3,996 (4.0%) were highly similar to a known plasmid sequence in NCBI, while 65,117 (64.7%) were novel sequences with no hits to any sequence in NCBI. As these different subsets of plasmids are partially overlapping, we took their union to find that 49,739 (49.4%) of predictions had some orthogonal support for being a plasmid, by MOB-suite or any of the first three types of analyses, and 28,658 (28.5%) had such support and were novel.

Novel plasmids are highly prevalent and reflect human biogeography, but their ecology is not explained by microbial taxonomy

Next, we sought to investigate the ecology of plasmids across human populations by creating a non-redundant collection of all plasmids and using metagenomic read recruitment to quantify their distribution across individuals. The dereplication step resulted in 11,121 non-redundant reference plasmids and 68,350 non-redundant predicted plasmids. Recruitment of short reads from the 1,782 globally distributed human gut metagenomes showed that only 1.9% of reference plasmids were present in at least two individuals in our dataset, revealing the limited ecological relevance of reference plasmids to naturally occurring gut microbial communities (Figure 3A). Such a weak detection of reference plasmids in the human gut is likely due to the heavy representation of human-gut associated plasmids in public databases that originate from a relatively small number of human pathogens that are not typically abundant in healthy humans. The reference plasmids did include those that were extremely prevalent across human metagenomes, such as pBI143, a cryptic plasmid that was present in 52% of the gut metagenomes in our dataset, which we investigated in-depth elsewhere (Fogarty et al. 2023). However, the predicted plasmids were much more prevalent across human populations in general (Figure 3B) where 63.1% of them were present in at least two individuals (Figure S2D). In fact, 99.7% of all plasmids that occurred in 100 or more individuals were predicted plasmids.

Due to their increased representation in naturally occurring gut microbiomes, plasmids we predicted from metagenomes better capture the biogeography and lifestyles of human populations compared to reference plasmids. Indeed, the organization of metagenomes based on their plasmid content showed that only 50.2% of individuals were placed next to someone from the same country based on reference plasmids (Figure 3A). This percentage increased to 74.0% with plasmids from metagenomes (Figure 3B). Furthermore, their distribution distinguished individuals from industrialized versus non-industrialized countries (Figure 3C), and revealed country-specific clustering of metagenomes (Figure S6).

These results parallel other studies that found associations linking the gut microbial taxonomy with the geography and lifestyles of human populations[56–59], and thus they lead to an important question: if plasmids and microbial taxa are each correlated with human geography, is microbial

taxonomy also correlated with plasmid distribution patterns? On one hand, it would be conceivable to expect a strong correlation between the two as plasmids rely on host machinery for replication and thus their presence in an environment depends on the presence of a suitable microbial taxon. On the other hand, such associations may be weak, or even nonexistent for two reasons: (1) some plasmids are known to have a broad host range[60–62], and thus their presence in a given environment might not be consistently associated with the presence of a single species or even higher taxonomic category such as a genus or phylum, and (2) plasmids can be gained or lost as a function of environmental pressures, such that nearly identical microbes can differ by the presence or absence of a plasmid or in the number of plasmid copies. By taking advantage of a large number of plasmids that are representative of human biogeography, we examined the ecological associations between plasmid distribution patterns and microbial taxonomy to determine to what extent plasmid ecology can be explained by microbial taxonomy.

For every plasmid, we inferred its most likely host as the taxonomic group that had the most similar ecological distribution (see Methods). While some predicted plasmids had a high ecological similarity with their best matching taxonomic group, the vast majority of predicted plasmids had low similarity scores (median correlation = 0.04, median Jaccard = 0.21) (see Methods, Figures S7A and S7B). We also observed low similarity scores even for reference plasmids that are isolated from a defined microbial host (Figures S7C and S&D). For example, the plasmid pDOJH10S and its cognate host, *Bifidobacterium longum*, were present together in 10 metagenomes; however, in 27 metagenomes we only found the plasmid, and in 69 metagenomes we only found the host (Figure S7E). A more careful detection of these sequences using read recruitment also showed low overlap in metagenomes, with a Jaccard index of 0.40 (Figure S7F, Table S13). The weak correlations between plasmid distribution patterns and microbial taxonomy suggest that plasmids are a highly complex and dynamic feature of microbiomes (Figures 3A, 3B, and S2D), forming an ecological dimension that can stratify human populations (Figures 3C and S6) in ways that cannot be explained by microbial taxonomy alone (Figure S7). While high-throughput analyses of human gut microbiomes often focus on taxonomic features, it has been challenging to find significant or reproducible taxonomic associations that distinguish health and disease states[63,64]. As plasmids often carry key determinants for survival in an environment, a complete understanding of the microbial ecology of health and disease states likely requires the inclusion of insights into plasmid ecology, which requires not only the recovery of naturally occurring plasmids through strategies such as PlasX, but also the

305 characterization of their gene pool in an evolutionary framework, which we aimed to do next using MobMess.

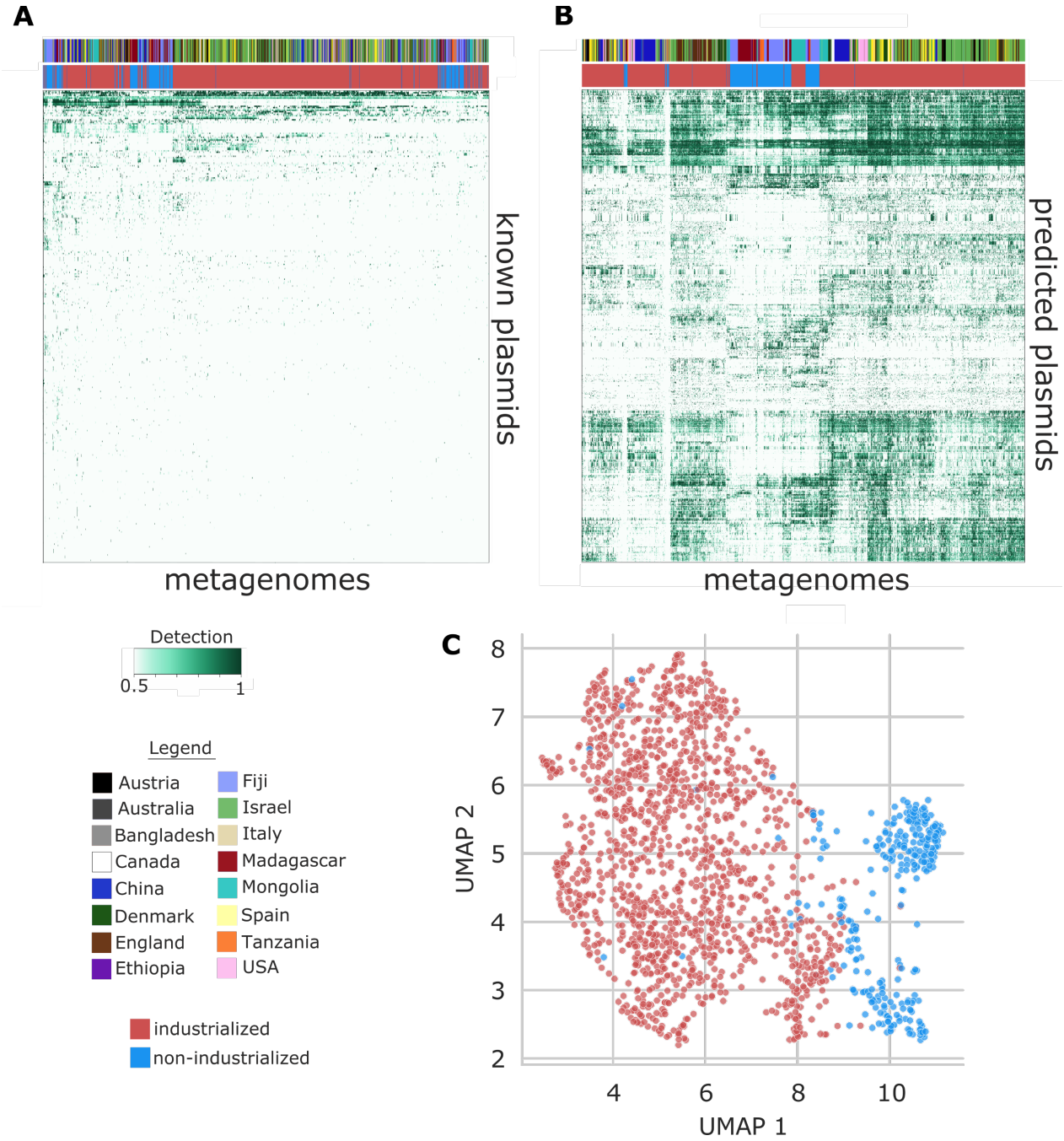


Figure 3. Global plasmid ecology. (A) Read recruitment of human gut metagenomes to 11,121 non-redundant reference plasmids. The heatmap shows the 338 plasmids that are present in at least one metagenome (≥0.95 detection). **(B)** Read recruitment to 68,350 non-redundant predicted plasmids. The heatmap shows the 1,000 most prevalent plasmids that are present in at least one metagenome and have PlasX score ≥0.75. In A and B, column colors indicate country of origin and lifestyle (industrialized or non-industrialized). **(C)** Clustering of metagenomes based on the predicted plasmids that are present, using the UMAP dimensionality reduction method[65]. Metagenomes from industrialized or non-industrialized populations are colored red or blue, respectively.

Plasmid systems organize evolutionarily related plasmids by distinguishing backbone versus cargo content

Our large collection of naturally occurring plasmids provides a unique opportunity to study evolutionary patterns in the human gut plasmidome. Due to frequent genetic rearrangements, a hallmark of plasmid evolution is the reuse of a backbone complemented with variable cargo/accessory genes[13,15,17,18]. Plasmid backbones, which so far have been characterized based on nucleotide identity[13,14,68], gene similarity[18], or gene annotations[69–71], typically encode machinery necessary for plasmid maintenance, while the cargo genes represent additional genetic content, such as antibiotic resistance or other fitness-determining functions.

Here we designed a novel network partitioning algorithm, MobMess (Mobile Element Systems), to study backbone structures in any collection of plasmid sequences at scale (Figure S8). Briefly, MobMess first calculates pairwise alignments across all plasmids to build an initial sequence similarity network, in which a directed edge represents the containment of one plasmid within another, defined by ≥90% sequence identity and ≥90% coverage of the smaller plasmid. We found that these thresholds represent a natural divide between related versus distant plasmids, as we observed a “valley” at these thresholds in the distribution of pairwise average nucleotide identities between all predicted plasmids (see Supplementary Information, Figure S9). Next, MobMess recognizes and collapses redundancy between plasmids, and analyzes patterns of connectivity in the network to identify ‘backbone plasmids’ that satisfy two criteria: (1) the backbone plasmid must be a circular element, inferred here by paired-end orientation (Figure 2B), to ensure that it is not an assembly fragment and can replicate as an independent element, and (2) a backbone plasmid must be found as a subsequence within one or more ‘compound plasmids’. These compound plasmids are composed of the backbone and additional cargo, indicating the ability to acquire or lose genes. Here we define a backbone and its compound plasmids as an evolutionary unit called a ‘plasmid system’ (Figure 4A).

MobMess differs in multiple ways from recently described clustering approaches that have been applied to plasmid sequences[72,73], which have only been tested on reference plasmids that are complete. In comparison, we designed MobMess to handle complex scenarios by distinguishing between fragmented versus complete (circular) plasmids, a strategy that renders MobMess more suitable to work with complex datasets such as metagenomic assemblies where sequence fragmentation is a common problem. Tracking the containment of smaller plasmids within larger

ones is also a significant computational challenge that is overlooked by previous approaches (Supplementary Information). MobMess can distinguish whether a group of plasmids simply share a level of homology through partial alignments or a more nuanced evolutionary relationship through a common backbone that can replicate independently (Figure S10). This way, MobMess minimizes spurious connections between distinct plasmid entities while maximizing its inference of evolutionarily cohesive groups as plasmid systems (Figure S11, Supplementary Information).

This definition of plasmid systems facilitates analyses of plasmid backbone versus cargo content as well as their ecology when used in conjunction with metagenomic data, much in the same way that pangenomes enable studies of core versus accessory gene content in microbial genomes. However, plasmid systems are a specific case of pangenomics, as it is unlikely to find a naturally occurring microbial genome composed only of core genes. In contrast, backbone plasmids represent a minimal entity that can propagate using only backbone genes. With its algorithmic considerations, MobMess provides an automated framework and vocabulary to explore the concept of plasmid systems across different studies and datasets.

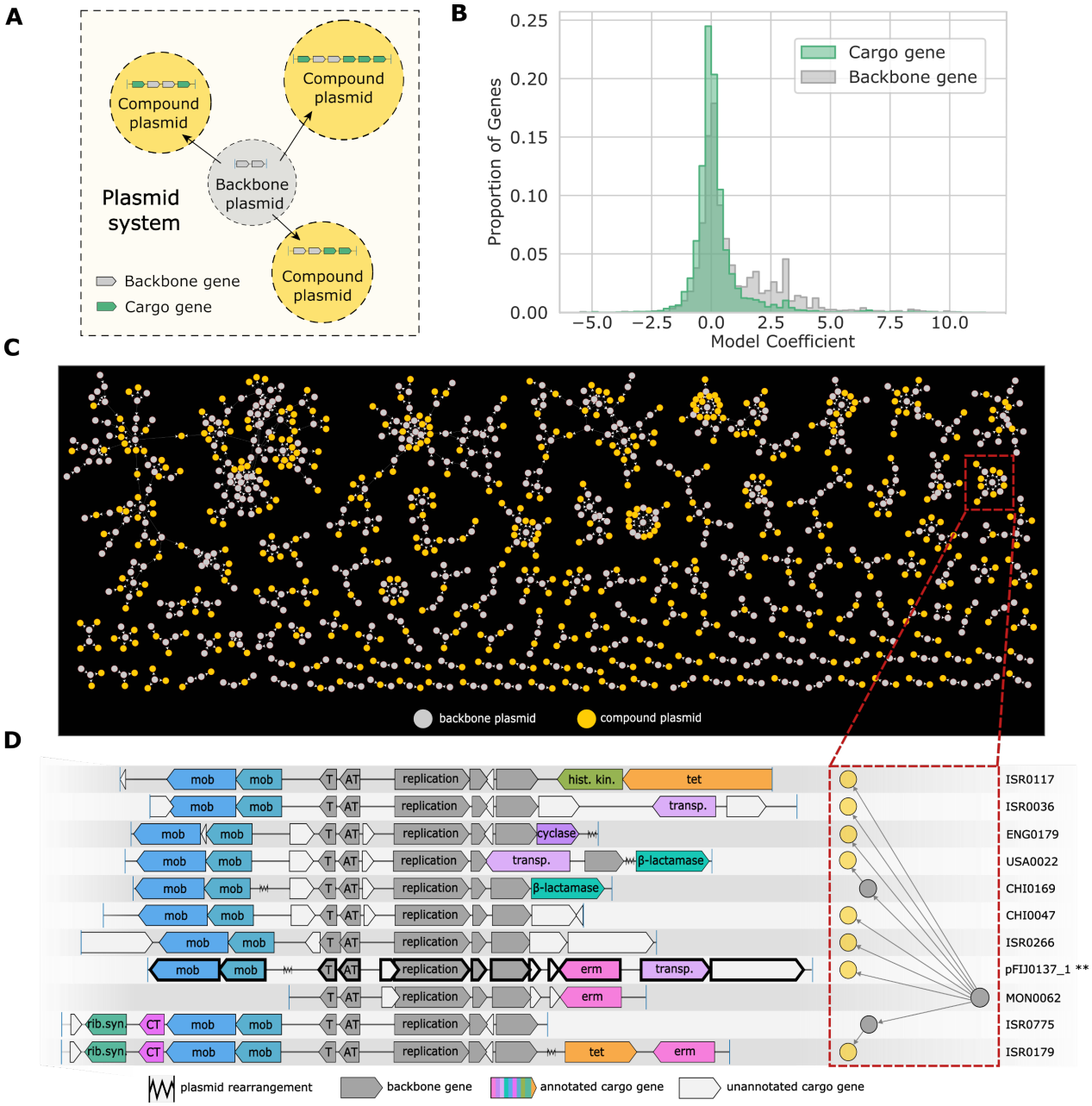


Figure 4. **Identification of plasmid systems. (A)** Network diagram of a plasmid system. **(B)** Distribution of model coefficients for backbone vs. cargo genes in the non-redundant set of 68,350 predicted plasmids. We excluded genes that lacked gene family annotations and thus have a coefficient of zero by default. We also excluded genes that were labeled as backbone with respect to some systems but cargo in others. **(C)** Network of all plasmid systems that contain ≥3 non-redundant and high-confidence plasmids. Only these types of plasmids are shown. **(D)** Genetic architecture of plasmids in PS486, encased by a red box in C. Two plasmids in C are excluded. The system's backbone (assembled from metagenome MON0062) encodes 5 backbone genes (colored gray). Rib.syn.=riboflavin biosynthesis, CT=conjugative transfer, mob=mobilization, T=toxin, AT=anti-toxin, tet=tetracycline resistance, erm=erythromycin resistance, transp.=transposon, hist. kin.=histidine kinase.

MobMess identifies 1,169 plasmid systems in human gut metagenomes with a wide repertoire of cargo functions

Our application of MobMess to the plasmid sequences predicted by PlasX from human gut metagenomes resulted in a total of 1,169 plasmid systems. Plasmid systems captured a small fraction of the genetic diversity among non-redundant plasmids (6.5%, or 4,424/68,350), however, they captured a large fraction of all circular plasmid contigs (72.7%, or 14,285/19,652) (see Methods, Table S9), and plasmids that were part of a system tended to be longer than plasmids that were *not* part of any system (Table S10). Due to our stringent criteria for the inclusion of plasmids in plasmids systems, MobMess identifies reliable plasmids with multiple representatives in a given dataset, independently of the initial confidence scores assigned by PlasX. For instance, of all plasmids with scores between 0.5 and 0.9, MobMess placed 16,663 in plasmid systems, which retrospectively suggests that a strict cutoff on PlasX prediction scores (such as >0.9) will remove many genuine plasmids from downstream analyses.

Plasmid systems were highly heterogeneous in their genetic complexity. 37 plasmid systems contained sequences that could be classified among 7 different plasmid incompatibility types (Inc11, Inc18, IncFIB, IncFIC, IncI-gamma/K1, IncK2/Z, IncW) (Table S9). 602 plasmid systems contained at least 2 non-redundant compound plasmids, with the largest system containing 168 non-redundant compound plasmids (Figure 4C). For example, pFIJ1037_1, the plasmid we isolated and transferred between *B. fragilis* organisms, was part of PS486, a system containing 24 non-redundant plasmids and found across a total of 127 metagenomes. PS486's backbone consists of a replication protein and a toxin-antitoxin system, and the cargo genes include beta-lactamases, erythromycin resistance, tetracycline resistance and riboflavin biosynthesis (Figure 4D, Table S11).

To understand how much genetic content is typically conserved or variable in a plasmid system, we calculated the percentage of genes on compound plasmids that were backbone genes versus cargo genes (see Methods). Plasmid systems spanned a wide range of cargo gene percentages between 0% and 100%, with a median value of 40% (Figure S12). Conversely, the median backbone percentage was 60%. PlasX often assigned higher model coefficients to backbone genes in the non-redundant set of predicted plasmids, suggesting these genes define the 'essence' of a plasmid by encoding essential functions that promote the ability of a plasmid to exist as a distinct element from the chromosome, such as the genes for plasmid replication, *repA*

(PF01051), and mobilization, *mobA* (PF03432) (Figure 4B). In contrast, PlasX assigned lower coefficients to cargo genes, suggesting they encode functions that are not universally essential but important for specific niches, such as nitrogen reductase, *nifH* (PF00142), and membrane transport, *ompA* (PF00691). Indeed, 24.1% (2,169/8,995) of backbone genes versus 13.4% (3,229/24,168) of cargo genes encoded COG and Pfam functions with descriptions related to plasmid replication, transfer, and maintenance (Supplementary Information).

The most frequent type of function encoded on cargo genes was antibiotic resistance, including efflux pumps, which can provide general resistance to multiple antibiotics, and genes targeting specific classes of antibiotics, such as glycopeptides and beta-lactams (Figure 5A). This large-scale observation is consistent with numerous examples of known plasmids encoding resistance and further illustrates how the widespread presence of these plasmids pose a public health threat[76–79]. Other highly prevalent cargo functions included a wide diversity of cellular and metabolic pathways defined in the COG (Figure 5B) and KEGG databases (Figure S13). The most enriched among these was tRNA modification, encoded in 35 compound plasmids within different systems. For example, the globally prevalent system PS1110 (present in 739 metagenomes) contained 291 compound plasmids (27 non-redundant), three of which encoded an enzyme that performs tRNA Gm18 2'-O-methylation (COG0566) and were collectively present in 498 metagenomes (Figure S14, Table S11). This enzyme is thought to reduce the immuno-stimulatory nature of bacterial tRNA, which is detected by Toll-like receptors (TLR7) of the mammalian innate immune system[80,81]. While plasmids in some pathogens are known to facilitate bacterial evasion of mammalian immune system by regulating surface proteins[82], the overwhelming prevalence of tRNA modification enzymes in our data suggests the likely presence of a previously unappreciated role for plasmids to increase the fitness of their bacterial hosts against the surveillance of the human immune system.

Overall, these results show that the plasmid systems we were able to identify with MobMess from the human gut represent evolutionarily cohesive units with the enrichment of different classes of functions in the backbone and cargo gene pools. Functional annotations further suggest that while the conserved pool of backbone genes can yield insights into plasmid compatibility and maintenance, the dynamic pool of cargo genes could serve as a means to identify genetic determinants of fitness that respond to particular environmental conditions.

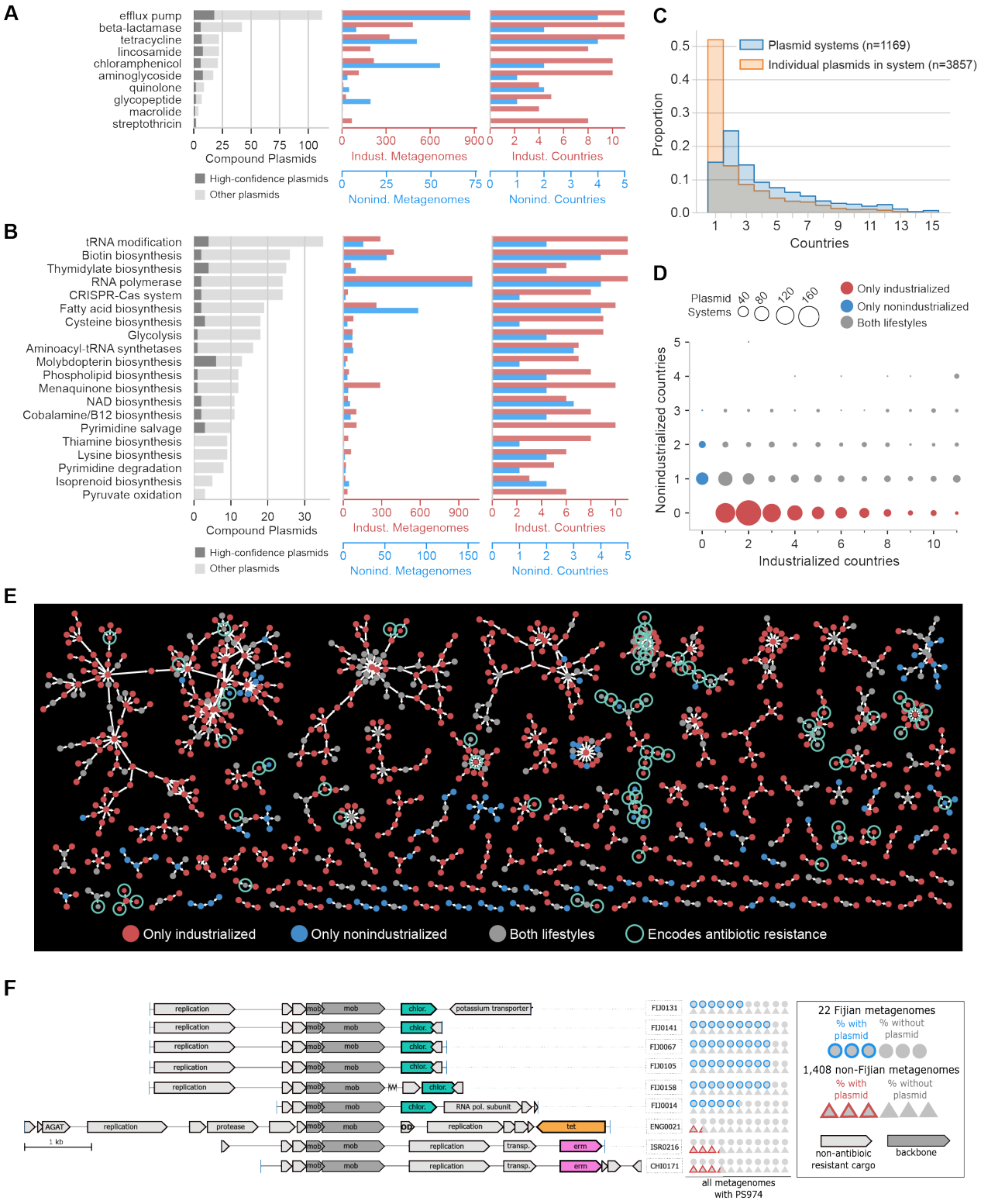
Plasmid systems reveal cargo genes adapted to specific environments

Given the highly heterogeneous biogeography of individual plasmids (Figure 3B) and their organization into their country of origin (Figure S6), we next asked whether the higher order evolutionary units described by plasmid systems simply consisted of plasmids with similar distribution patterns, or spanned larger geographical regions with individual plasmids of distinct ecology. Our analysis of plasmids and plasmid systems across metagenomes showed that while individual plasmids were often present in a single country, plasmid systems frequently spanned multiple countries (Figure 5C, Table S9). Indeed, of the 2,005 individual plasmids that were unique to a single country, 1,794 (89.5%) were part of more geographically diverse plasmid systems that were present in at least two countries. In fact, 84% (982/1,169) of plasmid systems in our dataset were present in at least two countries, and we found 9 plasmid systems were present in as many as 15 of the 16 countries (Table S9), suggesting that cargo genes that were likely selected for in different regions stemmed from conserved backbone structures.

The broad ecological distribution patterns of plasmid systems compared to the country-specific distribution patterns of plasmids they describe presents a unique opportunity to gain insights into environmental pressures that drive the composition of cargo genes within individual plasmid systems. To investigate this further, we focused on plasmid systems that spanned industrialized and non-industrialized countries, which were largely separated by the distribution of individual plasmids (Figure 3C). Many plasmids systems were exclusive to either industrialized or non-industrialized countries; however, 396 of them were present in both (Figure 5D). Antibiotic usage is a well-understood environmental pressure that often requires microbes to maintain plasmids with antibiotic resistance genes[83–90]. In our data, the evolution of antibiotic resistance in a plasmid system coincided with the ecological variation of compound plasmids in the system. Specifically, we identified 24 high-confidence, compound plasmids that encoded antibiotic resistance in cargo genes and were exclusively present in either non-industrialized or industrialized countries (Figure 5E). Among non-industrialized metagenomes, one of the most common types of antibiotic resistance is chloramphenicol resistance (Figure 5A). For instance, the plasmid system PS974 contained 97 non-redundant plasmids; however, this system possessed chloramphenicol resistance (conferred via an acetyltransferase) only in plasmids from Fiji (Figure 5F, Table S11). When we searched for these resistance plasmids across the global set of 1,430 metagenomes that contain PS974, we found them in 19/22 Fijian metagenomes but only in 1/1,408 non-Fijian

metagenomes (p=1.1 x 10^{-13}, Fisher's exact test) (Figure 5F, pictogram). Chloramphenicol is
480 routinely prescribed in Fiji to treat eye infections, central nervous system infections, periodontitis, shigellosis, typhoid and paratyphoid fevers, and diabetic foot infections, but it is rarely used in North America and Europe[91–94]. Strikingly, PS974 also contained compound plasmids that carry tetracycline resistance (171/1,408 metagenomes) or erythromycin resistance (429/1,408 metagenomes), yet these plasmids only occurred in individuals from China, Israel, and the United
485 Kingdom (Figure 5F). Matching sequences to these plasmids on the NCBI databases, suggested that their possible microbial hosts include populations in the phylum Firmicutes, such as *Blautia hydrogenotrophica*. However, the distribution patterns of plasmids in PS974 did not match any microbial taxa (highest Jaccard index=0.37 across all plasmid-taxon comparisons). Our observations here suggest that despite sharing a common backbone, compound plasmids in
490 PS974 give access to different antibiotic resistance genes, and their ecology is defined by lifestyle-specific usage of antibiotics.

While the connection between antibiotic usage and resistance is expected given previous studies[83–90], plasmid systems in general can be used to identify cargo genes that likely determine
495 fitness given known environmental pressures, as well as to hypothesize the existence of environmental pressures given known functions that differentially occur across environments. Overall, our results show that plasmid systems provide a biologically meaningful and computational framework to study plasmid ecology and evolution at scale.



500 **Figure 5. Functional and ecological variation of plasmid systems. (A-B)** The number of compound plasmids that encode antibiotic resistance (A) and COG pathways (B) in cargo genes. Also shown are the numbers of metagenomes and countries that contain those plasmids. In B, we show the 20 COG pathways

that have the highest number of compound plasmids. To avoid redundancy with A, we exclude COG pathways that occur in cargo genes encoding antibiotic resistance. **(C)** Prevalence of plasmid systems
505 versus the individual plasmids in those systems. **(D)** Distribution of plasmid systems based on the number of industrialized and non-industrialized countries they are found in. **(E)** Recoloring of the network of plasmid systems shown in Figure 4C. Colors indicate whether a plasmid occurred in only industrialized, only non-industrialized, or both types of countries. A green ring indicates a plasmid encoding antibiotic resistance. **(F)** Compound plasmids from PS974 that encode for resistance to chloramphenicol (chlor), tetracycline
510 (tet), or erythromycin (erm). 6/9 plasmids are circular. Dark gray genes are the backbone; light gray are cargo not related to antibiotic resistance. AGAT=aminoglycoside adenylyltransferase. OD=Oxaloacetate decarboxylase. PS974 is found in 22 Fijian and 1,408 non-Fijian metagenomes. The pictogram on the right-hand side represents these metagenomes using two shapes: circles (Fijian) and triangles (non-Fijian). For each plasmid, circles are colored blue to represent the proportion of the 22 Fijian-metagenomes that contain
515 the plasmid. Similarly, triangles are colored red to represent the proportion of the 1,408 metagenomes that contain the plasmid.

Discussion

Plasmids are found in nearly every microbial ecosystem, yet the computational challenges associated with their de novo identification have made it difficult for microbiologists to routinely survey plasmids and define evolutionarily cohesive units to describe plasmid diversity in complex environmental samples. PlasX and other plasmid recognition systems[30,31,35–37,45,95–97], along with MobMess to characterize plasmid systems, present a powerful roadmap for a detailed characterization of naturally occurring plasmids at scale. Even though our study focused on the human gut microbiome, we designed PlasX and MobMess using a broad collection of reference sequences, so that they can be applied to study any environment, and with the flexibility to include additional training sequences to improve accuracy. These methods provide a complementary approach to frequently used state-of-the-art workflows to study the taxonomic composition or functional potential of environmental or host-associated microbiomes through amplicon sequences or metagenomes.

Historically, plasmids and other genetic elements have been characterized on the basis of qualitative properties and descriptions. Early applications of machine learning approaches to predict plasmids employed a small number of marker genes, which limited the recognition of plasmids that lacked recognized markers. The effect of this shortcoming is clear, as many of the genuine plasmids contained no such markers (Figure 2C). Another popular strategy to recognize plasmids has been to employ k-mers, yet such algorithms will also miss plasmids that have low sequence similarity to those that are described in public databases (Figure 1D). By relying upon gene families, PlasX presents an unbiased addition to our bioinformatics toolkit to predict plasmids. But even though both our quantitative and qualitative surveys showed that PlasX surpasses the performance of state-of-the-art algorithms to identify plasmids, it is essential for researchers to take into consideration that predicted sequences will contain both false positives and false negatives. Plasmids can be difficult to distinguish from other mobile or integrated genetic elements as they share common features, including being extrachromosomal[32,98], facilitating horizontal gene transfer[99,100], or encoding traditional core functions like replication and mobilization[38,101]. PlasX's accuracy is also tied to the underlying training set of reference plasmids and chromosomes; over- or under-representation of certain types of sequences can bias the model and limit generalizability.

The confidence scores PlasX assigns to each prediction may serve as a means to adjust the amount of noise in prediction results; however, researchers must consider the trade-off between sensitivity (i.e., capturing a higher fraction of true plasmids) versus specificity (i.e.,

550 reducing the number of falsely predicted plasmids) while setting a cut-off. The trade-off between sensitivity and specificity was most visible in our cross-validation analyses, where a threshold of >0.5 lies at an inflection point in the precision-recall curve of Figure 1D (with a precision of 0.850 and recall of 0.500). This threshold was also sufficient at distinguishing ICEs and prophages from plasmids. Applying a stricter threshold of >0.9 increases precision by 13% (to 0.920), but it also
555 decreases recall by 44% (to 0.280). As our understanding of plasmid diversity in metagenomes is greatly underdeveloped, here we decided on applying a threshold of >0.5, in order to provide a reasonable balance between precision and recall for our study and to include many potentially novel plasmids in our results. A good example for this is the novel and long-missed Wolbachia plasmid[47], which has a score of 0.73. At the same time, stricter thresholds (such as >0.9) or other
560 filters such as circularity may be more appropriate in future work where a higher precision of predicted plasmids (fewer false positives) is a higher priority.

Given the dynamism of plasmids, organizing them into evolutionarily cohesive groups is a significant challenge. Previous computational methods for organizing plasmids have relied on average nucleotide identity (ANI) to represent the whole-sequence similarity between plasmids,
565 but while computational tractable, this single statistic misses other evolutionary dimensions that relate sequences. To address these issues we developed MobMess, which resolves the containment of plasmid sequences within one another. This methodological advance enabled us to identify plasmids systems, revealing the great extent to which plasmids in complex ecosystems are not static entities but actively evolving in response to the environment. Plasmid malleability is
570 a desirable property in bioengineering and has often motivated the repurposing of naturally occurring plasmids into major tools for genetically modifying organisms. In this vein, we propose computational identification of plasmid systems as an attractive approach to expand the toolkit of available plasmids for genetic engineering, particularly if they are found in isolates that currently lack tools to make them genetically tractable. Plasmid systems, which manifest in many distinct
575 forms across multiple human populations, excel at incorporating new functions and propagating across a wide range of natural environments and may behave similarly in laboratory settings.

An overarching implication of our findings is that high-throughput recognition and characterization of plasmids in microbiome studies are necessary for more complete insights into the ecology of naturally occurring microbial systems.

580

Methods

Compiling and annotating a reference set of plasmids and chromosomes. We obtained a list of 16,168 plasmids from the 2019_03_05 version of PLSDB[102]. We also downloaded the entire collection of 13,471 complete bacterial genome assemblies from NCBI RefSeq on October 26, 2019, using instructions at https://www.ncbi.nlm.nih.gov/genome/doc/ftpfaq/#allcomplete. The RefSeq assemblies contained 26,376 contigs, of which we discarded 11,350 that are also in PLSDB. The reference set of 16,827 plasmids consisted of 16,168 PLSDB contigs, as well as 659 contigs from the Refseq assemblies that were labeled as 'Plasmid' in the 'Assigned-Molecule-Location/Type' field of the NCBI assembly report. The reference set of chromosomes was the remaining 14,367 Refeq contigs.

To identify and annotate genes in these sequences, we used the program `anvi-run-workflow` with `--workflow contigs` implemented[103] in anvi'o[104] v7.1, which uses Snakemake[105] to execute previously defined steps (https://merenlab.org/anvio-workflows/) to generate anvi'o contigs-db files (https://anvio.org/m/contigs-db). These steps include first running Prodigal[106] to call genes and then running DIAMOND v2.0[107] and HMMER v3.3[108] on amino acid sequences to determine gene functions against the Cluster of Orthologous Groups of proteins (COGs)[42] and Protein Family Database models (Pfams) v32.0[43], respectively. To minimize noise, we used an e-value cutoff of 10^{-10} for COGs and the default model scores for Pfams.

Modeling *de novo* gene families. We inferred *de novo* gene families by running MMseqs2[109] v10.6d92c on all amino acid sequences in our reference plasmids and chromosomes. First, we ran `mmseqs clusthash` to collapse identical sequences into a non-redundant set for faster execution of the next step; the collapsing was inverted at the end to annotate all genes. Next, we ran `mmseqs cluster` to calculate pairwise alignments and then cluster genes that are aligned above a minimum sequence identity threshold (parameter `--min-seq-id`). We ran this program multiple times with different thresholds (0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.25, 0.2, 0.15, 0.1, 0.05) to infer a wide range of possible families. Families from different thresholds can be redundant, so we merged nested families, i.e. if family X contains all genes in family Y, then we keep X and discard Y. We also discarded any family that contains only one gene. In theory, families inferred from a higher threshold (e.g. 0.9) should always nest within a family inferred from a lower threshold (e.g. 0.05), such that we would discard all families from higher thresholds. But in practice, families don't always nest within each other but only overlap partially. After merging, our final model used the following number of families from each threshold.

Identity Threshold	# of De Novo Families
0.05	720,587
0.15	0
0.1	0
0.2	85,042
0.3	71,282
0.25	70,504
0.4	49,106

0.5	23,965
0.6	31,379
0.7	18,837
0.8	10,331
0.9	9,099

In total, our model used 1,090,132 gene families, which annotated 162,783,114 genes. Note that because these gene families can still overlap with each other, a gene may have multiple annotations. This analysis took advantage of MMseqs2's parallelism, taking ~6 hours using 256 CPU cores.

We refer to a *de novo* family as a 'subfamily' if 90% or more of its set of amino acid sequences are also annotated to a specific COG or Pfam. Note that this definition provides a small tolerance such that a subfamily does not need to be a perfect subset of a COG or Pfam. For the example about Pfam PF10609 (Figures 1F and 1G), we gathered the 253 amino acid sequences annotated to PF10609 and the subfamily mmseqs_5_1535552. We also gathered the 1,391 sequences annotated to PF10609 and the subfamily mmseqs_70_40217271. We collapsed 100% identical sequences to yield a total collection of 142 and 310 sequences from mmseqs_5_1535552 and mmseqs_70_40217271, respectively. We aligned all of these sequences together using muscle v3.8.1551 (default parameters)[110] and then constructed a maximum likelihood phylogenetic tree using IQ-TREE v2.1.2 (parameters -m TEST -bb 1000 -alrt 1000 -T AUTO)[111]. We then rooted the tree using the midpoint method.

Subtypes and slicing of reference sequences. To group reference sequences into subtypes, we used mash v2.2.2[112] (command `mash dist`, sketch size 100000, kmer size 21) to calculate a distance score of 0 to 1 between every pair of sequences. Next, we created an undirected graph, where sequences are nodes and sequences are connected if their distance is ≤0.1. We defined a 'subtype' as one of the 7,326 connected components in the graph. 3,935 subtypes contained only plasmids; 3,355 subtypes contained only chromosomes; and 36 subtypes contained both plasmids and chromosomes (Supplementary Table S1).

We sliced reference sequences into 10 kbp slices by sliding a 10 kbp window at 5kb increments. The first window starts at the beginning of the sequence, and the final window stops at the end of the sequence. For instance, a 23kb sequence would be sliced at 0-10 kbp, 5-15 kbp, 10-20 kbp, and 13-23 kbp. A slice was annotated with any gene that was entirely or partly inside the slice. In total, we generated 10,453,279 slices from the reference chromosomes and 343,246 slices from the reference plasmids.

Assessing model performance in cross-validation. We performed 4-fold cross validation by splitting the set of 10 kbp slices of reference sequences into four random subgroups. We used the sequences in three subgroups to train our model, PlasX, and then we evaluated the performance of the model on the fourth subgroup. We redo this procedure by changing which subgroups are used for training or for evaluation – for a total of four times. This procedure is a common technique in machine learning, more generally known as "k-fold validation" where k is the number of subgroupings. In a naive split, we keep all slices from the same reference

sequence together in either training or testing data. In an informed split, we keep all slices from the same subtype together.

We assigned weights to the 10 kbp slices when calculating precision and recall performance (Figure S1F and 1D). Consider the following notation to represent sequences:

$$S_i \;=\; sequence\ i$$
$$P_u \;=\; the\ set\ of\ plasmid\ sequences\ in\ subtype\ u$$
$$C_u \;=\; the\ set\ of\ chromosome\ sequences\ in\ subtype\ u$$
$$D_i^k \;=\; window\ slice\ k\ of\ sequence\ i$$

And consider the following notation to represent weights:

$$w(D_i^k) \;=\; weight\ of\ window\ slice\ k\ of\ sequence\ i$$

$$w(S_i) = \sum_k w(D_i^k) \;= weight\ of\ sequence\ i$$

$$w(P_u) \;=\; \sum_{S_i \in P_u} w(S_i) = weight\ of\ plasmid\ sequences\ in\ subtype\ u$$

$$w(C_u) = \sum_{S_i \in C_u} w(S_i) \;= weight\ of\ chromosome\ sequences\ in\ subtype\ u$$

We defined two different scenarios for assigning weights. Scenario A satisfies the following conditions:

1.) All slices from the same sequence have the same weight

$$w(D_i^s) = \; w(D_i^t) \quad \forall s, t$$

2.) The weight of every sequence is equal to 1

$$w(S_i) = 1 \quad \forall i$$

Scenario B satisfies the following conditions:

1.) All slices from the same sequence have the same weight

$$w(D_i^s) = \; w(D_i^t) \quad \forall s, t$$

2.) All plasmid (or chromosome) sequences in the same subtype have equal weight

$$w(S_i) \;=\; w(S_j) \quad \forall\ i, j, u\ where\ S_i \in P_u\ and\ S_j \in P_u$$
$$w(S_i) \;=\; w(S_j) \quad \forall\ i, j, u\ where\ S_i \in C_u\ and\ S_j \in C_u$$

3.) All subtypes have equal weight

$$w(P_u) = w(P_v) \quad \forall u, v$$
$$w(C_u) = w(C_v) \quad \forall u, v$$

4.) The sum of weights across all slices equals the total number of slices

$$\sum_u \quad w(P_u) = total\ number\ of\ plasmid\ slices\ (i.e. 343{,}246)$$

$$\sum_u \quad w(C_u) = total\ number\ of\ chromosome\ slices\ (i.e. 10{,}453{,}279)$$

Each scenario implies a unique assignment of weight values. Scenario A requires that every sequence has the same weight. Importantly, this ensures that long sequences, which have disproportionately more slices, have equal weight as shorter sequences. Scenario B further

requires that every subtype has the same weight. Importantly, this ensures that subtypes that contain a disproportionately large number of sequences (e.g. subtypes that represent commonly studied bacteria, such as *Escherichia*, *Salmonella,* and *Klebsiella*) have equal weight as subtypes with fewer sequences.

We evaluated performance under two different cross-validation and weighting scenarios. Figure S1F shows the result of training models using a 'naive' cross-validation split and calculating precision/recall using weights from Scenario A. Figure 1D shows the results of training models using an 'informed' cross-validation split and calculating precision/recall using weights from Scenario B. We calculated precision/recall using the function *sklearn.metrics.precision_recall_curve* from the scikit-learn Python package[113], with the parameter *sample_weight* set to the weights of the slices. We calculated AUCPR with the function *sklearn.metrics.average_precision_score*.

PlasX implementation. We implemented PlasX as a logistic regression using the *SGDClassifier* class from scikit-learn[113]. For training and evaluating PlasX, we used 10 kbp slices of the reference sequences, to normalize for the fact that chromosomes are generally much longer than plasmids and to improve downstream application of PlasX on sequence collections that may contain a large number of fragmented sequences, such as assembled metagenomes. Regardless of how we evaluated PlasX (either Figure S1F or Figure 1D), we always trained it with weights defined by Scenario B and based on only slices in the training data.

To improve performance, PlasX uses a technique called elastic net regularization, which identifies gene families with redundant or noisy signals and then minimizes the usage of these families by setting their coefficients equal or close to zero. Consequently, only a non-redundant and informative set of gene families can impact predictions by having coefficients far from zero (Figure S1D). To implement elastic net regularization, we performed a grid search of hyperparameters, with the regularization parameter *alpha* ranging from 10^{-8} to 10^{-3} in multiplicative increments of $\sqrt{10}$ and the parameter *l1_ratio* being 0.0, 0.25, 0.5, 0.75, or 1.0. For each evaluation scenario (either Figure S1F or Figure 1D), we selected the hyperparameters that produced the best performance. We used the best hyperparameters from the 'informed' cross-validation and the weights defined by Scenario B (*alpha*=3.16×10^{-6}, *l1_ratio*=0.0) to retrain PlasX on all 10 kbp slices and create the final model that we used to predict plasmids from metagenomes.

Predicting plasmids from metagenomic assemblies. We downloaded fastq files for 1,782 short-read and paired-end metagenomes from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) using the program `fastq-dump`. The countries represented are Austria[114], Australia (https://www.ncbi.nlm.nih.gov/bioproject/PRJEB6092), Bangladesh[115], Canada[116], China[117,118], Denmark[119], England[120], Ethiopia[58], Fiji[121], Israel[122], Italy[123], Madagascar[58], Mongolia[124], Spain[125], Tanzania[123] and USA[126,127]. Some samples were sequenced multiple times (i.e. multiple records in SRA), in which case we concatenated the fastq files together. We have separated these multiple accessions using the delimiter '|'. We labeled Tanzania, Ethiopia, Bangladesh, Madagascar, and Fiji as non-industrialized and the other countries as industrialized.

All steps of quality filtering, metagenomic assembly, read recruitment and profiling were automated using snakemake[105] workflows in anvi'o[128]. The `illumina-utils`[129] commands `iu-gen-configs` and `iu-filter-quality-minoche` with the flag `--ignore-deflines` were used to quality filter the raw paired-end reads. Each metagenome was assembled individually using IDBA_UD[130] with default settings, except the flag `--min_contig 1000`.

We annotated COGs and Pfams in all assembled contigs using the same procedure as the reference plasmids and chromosomes. To annotate *de novo* families, we first used `mmseqs result2profile` (default parameters) to represent the sequence conservation in each *de novo* family as a profile. We then used `mmseqs search` (default parameters) to search for profiles across all genes. We kept hits where the alignment coverage was ≥80% of both the gene and the profile and where the alignment identity was at least ≥X-0.05 where X is the minimum identity threshold used to originally construct the family (parameter --min-seq-id). For example, if a family was constructed using an identity threshold of 0.8, then we kept hits with an identity ≥0.75. Using these gene annotations, we ran PlasX to assign a score to every contig. We kept contigs intact, rather than slicing them into 10 kbp windows. Contigs with score >0.5 were classified as plasmids.

Detection and circularity of plasmids across metagenomes. We recruited short reads from our collection of metagenomes using Bowtie2 v.2.0.5[131]. We used the snakemake workflows in anvi'o to automate execution of bowtie and post-processing to calculate 'detection', i.e. the proportion of a sequence that is covered by at least one read. We ran bowtie2 using the three following combinations of parameters and input files.

First, to identify circular contigs, we recruited each metagenome's reads to a fasta file that contained only the contigs assembled from that metagenome. For computational efficiency, we ran bowtie2 with its default behavior to align every read at most once. We then analyzed the orientation of paired-end reads (Figure 2B). During assembly, circular sequences are broken by an artificial breakpoint to represent them as linear contigs. Consequently, DNA sequencing that occurred across this breakpoint will produce paired-end reads that align in a reverse-forward orientation to the ends of the contig. In contrast, if a sequence is not circular, then all paired-end reads are expected to align in a forward-reverse orientation. To illustrate this intuition, suppose the upstream read of a paired-end maps to positions 200-300 of a contig and the downstream read maps to 500-600. If the upstream read maps with a reverse complement strandedness (i.e. 'reverse') and the downstream read maps with the same strandedness as the way the contig is written (i.e. 'forward'), then the paired-end is in a reverse-forward orientation. In other words, if the contig is written 5'-to-3', then the upstream read maps 3'-to-5' and the downstream read maps 5'-to-3'. Inversely, the paired-end is in a forward-reverse orientation if the upstream read maps 5'-to-3' and the downstream read maps 3'-to-5'. Next, we defined the gap (or insert) size of a paired-end to be the distance between the closest (or farthest) aligned positions between its two reads. In our example, the gap size is 600-200=400 and the insert size is 500-300=200. Let *D* be the contig's length minus three times the median insert size of all forward-reverse paired-ends that are aligned to the contig. Finally, we label a contig as circular if (1) its detection was ≥0.95 and (2) it had at least one reverse-forward paired-end with a gap greater than or equal to *D*. This approach of examining reverse-forward paired-ends was inspired by [132]. There were 154,680 contigs that were not predicted to be plasmids but still appeared circular; however, these contigs tended to have a smaller number of supporting reverse-forward reads relative to their coverage

(Figure S2E), which may indicate that they are other types of mobile elements such as viruses or ICEs that temporarily circularize.

Second, to study the ecological distribution of all plasmids and plasmid systems at the same time, we recruited each metagenome's reads to a fasta file that contained either the non-redundant set of 68,350 predicted plasmids or the non-redundant set of 11,121 reference plasmids. For computational efficiency, every read was aligned at most once (i.e. the default behavior of bowtie). We designated a plasmid as being present in a metagenome if its detection was ≥0.95. To compare metagenomes based on their plasmid content in Figure 3 and Figure S6, we ran UMAP v0.5.1[65] with parameters `n_neighbors=30, n_components=2, min_dist=0.15, metric='jaccard', random_state=1`. The heatmaps in Figure 3 were generated using the `heatmap.2` package in R, with agglomerative clustering using median linkage on Euclidean distances.

Third, to study the specific plasmids from PS974 and PS1110 that were shown in Figure 5F and Figure S14 (see Table S11 for contig names), we ran bowtie2 on each sequence separately. This setup allowed every read to align to potentially multiple sequences, resulting in a more complete estimation of which metagenomes contained a plasmid. For the backbone sequences of these systems, we designated them as present in a metagenome if their detection was ≥0.95. For compound plasmids in PS1110 that encoded a chloramphenicol resistance gene, we designated them as present in a metagenome if they satisfied an additional criterion that ≥0.95 of the resistance gene was covered by at least one read.

Estimation of potential hosts for plasmids in metagenomes. We used two different formulas to calculate the ecological similarity between a plasmid and potential host: (1) the Pearson correlation in the abundance levels of the plasmid and host across metagenomes, and (2) a Jaccard index to represent the fraction of metagenomes that contain both the plasmid and host.

We estimated taxonomic abundances in every metagenome by running kraken2[133] v2.1.2 with its standard database (https://github.com/DerrickWood/kraken2) and then refined the abundances using bracken[67] v2.5 (https://github.com/jenniferlu717/Bracken) with database parameters `-k 35 -l 96`. We ran bracken (parameter `-r 96`) a separate time for every taxonomic rank: S1 (subspecies/strain), S (species), G (genus), F (family), O (order), C (class), P (phylum), D (domain). The output of this analysis is a count of how many reads originated from each taxon.

To compare the metagenomic presence/absence of plasmids versus taxa, we calculated M_P, the set of metagenomes where a plasmid *P* is detected at ≥95% (based on recruitment to the 68,350 non-redundant plasmids), and M_T^r, the set of metagenomes in which at least *r* reads originated from taxon *T* (based on kraken and bracken). For each plasmid, we attempted to find the best explanation of its ecological distribution by comparing the plasmid to every taxon using the Jaccard index, and by scanning many possible read thresholds. More exactly, we used the following formula to represent the best possible explanation of a plasmid's ecological distribution:

$$\max_T \max_r Jaccard(P, T; r)$$

where

$$Jaccard(P, T; r) = \frac{|M_P \cap M_T^r|}{|M_P \cup M_T^r|}$$

We evaluated 29 values for the threshold *r*, ranging from 1 read to 10 million reads in multiplicative
830 increments of $10^{1/4}$. We ignored plasmids that were present in less than 5 metagenomes, i.e. $|M_P|<5$, because it was likely that these plasmids would have a high Jaccard similarity to some taxon by random chance. For instance, we observed that many pairs of plasmids and taxa occur in exactly one and the same metagenome, and thus they have a Jaccard index of 1.

To compare continuous-valued abundances, we defined a plasmid's abundance in a
835 metagenome as the sum of coverage values across all sequence positions divided by sequence length, and we defined a taxon's abundance as bracken's estimate of the number of reads originating from the taxon. If a plasmid had detection of <95%, then we set its abundance to 0. If a taxon had less than 1000 reads, then we set its abundance to 0. We ignored plasmids and taxa that had non-zero abundances in less than 5 metagenomes. For every pair of plasmid and taxon,
840 we estimated the Pearson correlation between their abundance levels across metagenomes using FastSpar[134] v1.0.0 (https://github.com/scwatts/fastspar), which is an improved implementation of the SparCC[66]. This method accounts for the compositional nature of the data—in which abundances reflect relative instead of absolute quantities—by assuming that the amount of correlations in a data is sparse. We ran FastSpar on the non-redundant set of predicted
845 plasmids, and ran it separately on the non-redundant set of reference plasmids.

We performed a more careful analysis of plasmid pDOJH10S and its cognate host *B. longum DJO10A* by performing a read recruitment of metagenomes to both sequences together, using bowtie2 and allowing every read to align at most once (Supplementary Figure S7F, Table S13). Following Utter et al.[135], we applied a >50% detection threshold to identify the presence of
850 the plasmid or host. 365 metagenomes contained the plasmid, 818 metagenomes contained the cognate host, and 336 metagenomes contained both genomes (Jaccard index = 0.40).

Keyword analysis of COGs and Pfams for plasmid functions

We labeled COGs and Pfams as being a plasmid-associated function (Figure 1E) if its database
855 description contains any of following keywords as a substring: *'plasmid', 'toxin', 'replicat', 'integrase', 'transpos', 'recombinase', 'resolvase', 'relaxase', 'recombination', 'partitioning', 'mobilis', 'mobiliz', 'type iv', 'conjugal', 'conjugat', 'segregat', 'MobA', 'ParA', 'ParB', 'BcsQ'*. We labeled backbone and cargo genes as being related to plasmid replication, transfer, or maintenance if they were annotated to any plasmid-associated COG or Pfam (see section
860 "Classification of cargo and backbone genes").

To determine if a predicted plasmid is 'keyword-recognizable' (Figure 2C), we searched the plasmids for COGs and Pfams using a more restricted set of keywords (just "plasmid" and "conjugation") instead of the keywords above.

865 **MobMess algorithm to dereplicate plasmids, remove assembly fragments, and discover plasmid systems.** The MobMess algorithm performs three tasks. It de-replicates plasmids that are nearly redundant to each other; it removes plasmids that appear to be assembly fragments; and finally it organizes plasmids together into evolutionary groups called plasmid systems. MobMess consists of several steps described below.

870 MobMess first performs an all-vs-all pairwise alignment of sequences using the MUMmer alignment package (v4.0.0rc1)[136]. All sequences are placed into a single fasta file and then aligned with `nucmer` (parameters `--maxmatch --minmatch=16`) to calculate local alignment

blocks. Alignments are specified asymmetrically such that one sequence is designated as the query *q* and the other is the reference *r*. For every *q* and *r*, the alignment blocks calculated by `nucmer` are written to a separate file, and then a subset of blocks is identified using `delta-filter` (parameters `-q -r`) to create a one-to-one alignment.

Next, MobMess constructs a directed graph *G* where vertices are sequences and edges represent the containment of one sequence within another (Figure S8A). Formally, consider a query *q* and reference *r*. Let |*q*| be the length of *q*. For the ith alignment block between *q* and *r*, let s^i , e^i , and δ^i be the start position in *q*, end position in *q*, and number of alignment mismatches and indels, respectively. The following values summarize the information across all alignment blocks between *q* and *r*.

Sum of block lengths $$L = \sum_i e^i - s^i$$

Number of mismatches and indels $$E = \sum_i \delta^i$$

Proportion of query positions covered $$C = \sum_{j=1}^{|q|} \begin{cases} 1/|q| & \text{if } \exists i \text{ such that } s^i \leq j \leq e^i \\ 0 & \text{otherwise} \end{cases}$$

Local sequence identity $$I_{local} = (L - E)/L$$

Global sequence identity $$I_{global} = I_{local} * C$$

MobMess creates a directed edge (*q*,*r*) in *G* if I_{local} and *C* are above user-specified thresholds. In this study, we applied thresholds of $I_{local} \geq 0.9$ and $C \geq 0.9$ (see Supplementary Information). In Figure S9, we re-ran MobMess using various thresholds on I_{local} and *C* (the same threshold was applied to I_{local} and *C* at the same time).

MobMess clusters sequences according to strongly connected components in *G*, calculated with igraph v0.8.2[137] in Python. That is, two sequences *x* and *y* are placed in the same cluster if there exists a directed path from *x* to *y* and another from *y* to *x* in *G*. Intuitively, a cluster represents a set of sequences that are nearly identical to each other across nearly their entire lengths. MobMess then reduces *G* to another graph *H*, called the condensation graph, by contracting every cluster of sequences into a single vertex. A directed edge (*u*,*v*) exists in *H* if and only if there are sequences $x \in u$ and $y \in v$ where edge *(x,y)* exists in *G*. Note that *H* does not have any cycles. As proof by contradiction, if there were a cycle of clusters, then those clusters would have been in the same strongly connected component in *G* and hence would have been merged into a single, larger cluster.

MobMess labels every cluster in *H* as one of the three following types: (1) a 'backbone cluster' if it has an outgoing edge and at least one of its member sequences is circular, (2) a 'fragment cluster' if it has an outgoing edge but none of its member sequences are circular, or (3) a 'maximal cluster' if it does not have any outgoing edges. Intuitively, a maximal cluster represents the longest version of a plasmid observed in the data. In contrast, a backbone or fragment cluster represents a set of plasmids that are subsequences of other plasmids in a maximal cluster. The only difference between backbone and fragment clusters is that backbone clusters contain at least

one circular plasmid (implying complete assembly), while fragment clusters do not contain any
910 circular plasmids (suggesting they are assembly fragments of the maximal cluster).

To dereplicate sequences, MobMess discards all fragment clusters and then chooses a representative sequence from every maximal and backbone cluster. A cluster's representative is the sequence with the highest global sequence identity (I_{global}), averaged across the set of alignments where that sequence is the reference and other sequences in the same cluster are
915 the queries.

MobMess defines a plasmid system as a specific backbone cluster together with its 'compound' clusters, which are the set of non-fragment clusters connected to the backbone in *H*. Thus, there is a one-to-one correspondence between backbone clusters and plasmid systems. Note that systems can be nested within each other, because backbone clusters can be connected
920 to each other in *H*. Thus, a backbone cluster can be *the backbone* that forms a given plasmid system, and at the same time, it can also be a *compound cluster* with respect to an even smaller backbone that forms a different system. As another note, a maximal cluster can be a 'compound' cluster of a system, but it is also possible that some maximal clusters are not found in any system because they are not connected to any backbone clusters in *H*.

925 We ran MobMess to analyze the 226,194 predicted plasmid contigs. MobMess grouped the contigs into a total of 132,616 clusters. 64,266 clusters were 'fragment clusters' that contained 125,475 contigs, which we interpreted as assembly fragments of other predicted plasmids. We discarded these fragments from further analysis. The other 68,350 clusters were non-fragment clusters (i.e. 1,169 backbone and 67,181 maximal clusters) and contained 100,719 contigs.
930 Finally, MobMess identified 1,169 plasmid systems, which together represent 1,169 backbone and 63,926 maximal clusters (3,255 maximal clusters were excluded). See Figure S3 for a diagram of these numbers.

We ran MobMess separately on the 16,827 reference plasmid sequences, yielding 11,121 clusters. We assumed that all reference plasmids were circular, and thus there were no fragment
935 clusters. We visualized networks with Cytoscape[74] v3.8 and laid nodes out using the prefuse directed force layout[75]. While we have focused on plasmids, MobMess could be applied to dereplicate and organize other mobile genetic elements into systems.

Classification of cargo and backbone genes. We classified all genes on the backbone
940 plasmids of a plasmid system as backbone genes. For genes on compound plasmids, we tested whether the genes shared any *de novo* family annotations with the genes on the backbone plasmids. If so, we classified those genes as backbone genes, otherwise as cargo genes. For this analysis, we used the 1,090,132 *de novo* families that we constructed from reference plasmids and chromosomes in order to train PlasX, and we also used an additional set of 439,584 *de novo*
945 families that we constructed by running the command MMseqs2 (--min-seq-id 0.05) on the genes from all plasmid sequences in this study (16,827 reference and 226,194 predicted plasmids). These additional families allowed us to capture gene families that might be absent in reference sequences but are conserved in predicted plasmids. Note that the classification of genes as backbone or cargo depends on which plasmid system is being considered. It is possible for a
950 gene to be classified as a backbone gene with respect to one plasmid system and, at the same time, as a cargo gene with respect to another system. This is because a plasmid can be a backbone plasmid of a system and also be a compound plasmid of a different system.

For every non-redundant compound plasmid in the system, we calculated the fraction of genes in the plasmid that were cargo genes. We then averaged this fraction across all non-redundant compound plasmids in the system to define the "cargo gene percentage" of the system (Figure S12). Because every gene is either backbone or cargo, the percentage of backbone genes is 100% minus the cargo gene percentage.

For Figure 4B and to analyze the content of backbone/cargo genes, we used a non-redundant and unambiguous set of 8,995 backbone and 24,168 cargo genes. To derive these sets of genes, we first considered the 47,172 genes encoded on the 4,424 non-redundant plasmids that were part of at least one plasmid system. Of these 47,172 genes, we used the 8,995 genes that were classified as backbone genes because they were encoded on a backbone plasmid and that were never classified as cargo genes in any plasmid system. 24.1% (2,169/8,995) of these genes had a plasmid-associated keyword in their COG/Pfam annotations (see Methods section "Keyword analysis of COGs and Pfams for plasmid functions"). We also used the 24,168 genes that were always classified as cargo genes and never backbone genes in any plasmid system. 13.4% (3,229/24,168) of these genes had a plasmid-associated keyword. We excluded from analysis the 1,917 genes that were sometimes classified as backbone genes and other times cargo genes, depending on the system. We also excluded 12,092 genes that were on compound plasmids but were classified as backbone genes, as these genes are redundant with the backbone genes that were encoded on the backbone plasmid.

Identification of antibiotic resistance genes. We annotated antibiotic resistance genes using two databases. First, we searched against a database of resistance protein family HMMs from Resfams[138] (v1.2, dated 2015-01-27, 'Core' database at http://www.dantaslab.org/resfams). We used `anvi-run-hmms` from anvi'o[104] to automate running `hmmsearch` from HMMER[108] 3.3.2 and apply an e-value cutoff of 10^{-10}. Second, we ran rgi (v5.2.0, https://github.com/arpcard/rgi) to search for similarity in the CARD database of resistance genes[139]. We removed CARD hits that were labeled as 'Loose' and kept those labeled as 'Perfect' or 'Strict'. We removed any Resfams or CARD hits that contained the keywords 'transcription', 'regulat', 'modulat' in their database description, to avoid cases (e.g. TetR protein) where the hit is a gene that regulates the expression of another resistance gene but doesn't itself perform the molecular process that confers resistance. We categorized hits into major antibiotic resistance classes by searching for the following keywords in their functional descriptions: 'lincosamide', 'macrolide', 'erythromycin', 'chloramphenicol', 'aminoglycoside', 'streptothricin', 'glycopeptide', 'efflux pump', 'beta-lactamase', 'nitroimidazole', 'tetraycyline', 'quinolone', 'sulfonamide'. Additionally, we searched the extra keywords 'Van' and 'VanZ' to identify glycopeptide resistance; 'efflux', 'permease', and 'pump' to identify efflux pumps; and 'TetX' to identify tetracycline resistance.

High molecular weight (HMW) DNA extraction, long-read sequencing, and determination of circularity through long-reads. We employed a long-read sequencing strategy on two *Bacteroides fragilis* cultivars from two patients (p-214 and n-216 previously described in Vineis et al[55]). We extracted total genomic HMW DNA by one of two methods. For *B. fragilis* p-214, we used the Qiagen Genomic Tip 20/G procedure (also known as Method #4/GT) as previously described[140] on a 10 mL overnight BHIS broth culture. For *B. fragilis* n-216, we used a Phenol Chloroform protocol on 25 mL overnight BHIS broth cultures. Libraries were prepared with the

Rapid Barcoding Kit (SQK-RBK004) and the standard protocols from Oxford Nanopore Technologies (ONT) with a few modifications. For *B. fragilis* p-214, DNA fragmentation was performed on 6 ug DNA using 5 passes through a 22G needle in a 30 µL volume. The gDNA input was 1.5 µg (Table S12), based on sample availability in a 7.5 µL volume, with 2.5 µL Fragmentation mix added. We sequenced for 72 hours using a single R9.4/FLO-MIN106 flow cell (ONT). For *B. fragilis* n-216, DNA fragmentation was performed on 10 ug DNA using 10 passes through a 22G needle in a 250 µL volume. The gDNA input was 0.32 to 0.44 µg, based on sample availability in an 8.5 µL volume, with 1.5 µL Fragmentation mix added per sample. We sequenced for 72 hours using a single R9.4/FLO-MIN106 flow cell. We used Guppy (v4.0.15) for all post-run base calling, sample de-multiplexing and the conversion of raw FAST5 to FASTQ files.

To determine circularity, we used BLAST to align the long reads with a minimum quality score of 7 to our predicted plasmid sequences following a previously described approach[47]. During assembly, all DNA short reads are assembled as linear sequences even if they are circular elements. Circular elements have an artificial breakpoint to represent them as linear sequences, and this breakpoint can happen anywhere on the sequence depending on the assembly method. We identified and manually confirmed 500 long reads that align completely to a plasmid but not to the host chromosome (Figures S4A and S5B). We tested for the presence of an artificially introduced breakpoint by visualizing these alignments on the sequence as if it were assumed to be a circular element (Figure S5A). If indeed the sequence is circular, the long reads would overlap each other and “wrap around” the entire circumference of the sequence. In other words, all nucleotide positions of the sequence would be covered by at least one read and there would also exist a read that spans the breakpoint by aligning to both sides of the breakpoint. This property ensures the breakpoint is artificial, and hence the sequence is a circular element. Inversely, this property does not hold when the breakpoint is not artificial (i.e. the sequence is actually an assembly fragment or linear element). Some of these long reads align across the artificial contig breakpoint, indicating these plasmids are extrachromosomal and circular.

Transfer of predicted plasmid between microbial populations. In duplicate, we streaked *B. fragilis* 214 (donor, carries erythromycin resistant on pFIJ0137_1, one of 14 isolates from Vineis et al. [55]) and *B. fragilis* 638R (recipient, rifampicin resistant) onto plates with brain-heart infusion agar supplemented with hemin and vitamin K (BHIS). We picked colonies and incubated them in 5 mL BHIS media anaerobically at 37℃ for 20 hours. While pFIJ0137_1 lacks conjugation machinery, it contains two relaxases (blue genes in Figure S4) and thus could be mobilized by different conjugative apparatus in the host cell. To mate the donor to the recipient, 250 µL of donor cells were pelleted in a centrifuge at 5,000x gravity. We discarded the supernatant and resuspended the donor in 1 mL of the recipient culture. Again, cells were pelleted at 5,000x gravity, then resuspended in 25 µL of BHIS media. Cells were spotted onto BHIS agar plates and incubated anaerobically for 24 hours. After incubation, cells were resuspended in 1 mL BHIS. 250 µL of this suspension was plated onto BHIS plates containing 8 µg/mL rifampicin and 25 µg/mL erythromycin to select for *B. fragilis* 638R recipients of pFIJ0137_1. Duplicate plates each had approximately 300 colonies. Plating the donor or recipient alone on rifampicin erythromycin plates resulted in zero colonies, confirming the transformants were not spontaneous mutants to either antibiotic. Two transformant colonies were restreaked onto fresh BHIS plates containing 8 ug/mL rifampicin and 25 µg/mL erythromycin. Through short-read sequencing of the donor, recipient,

and resulting transconjugants, and by employing a read recruitment analysis, we confirmed that pFIJ0137_1 transferred from B. fragilis 214 to B. fragilis 638R (Figure S4).

Short-read sequencing of isolate genomes and confirmation of plasmid transfer. We grew 20-hour cultures of *B. fragilis* 214 donor, naive *B. fragilis* 638R, and *B. fragilis* 638R transconjugants containing pFIJ0137_1. Using the QIAseq FX DNA library kit (Qiagen), libraries of these strains were prepared with 100 ng of genomic DNA. DNA was fragmented enzymatically into smaller fragments and desired insert size was achieved by adjusting fragmentation conditions. Fragmented DNA was end repaired and 'A's were added to the 3' ends to stage inserts for ligation. During the ligation step, Illumina compatible Unique Dual Index (UDI) adapters were added to the inserts and the prepared library was PCR amplified. Amplified libraries were cleaned up, and QC was performed using a tapestation. Libraries were sequenced on Illumina MiSeq platform using v2 cassette to generate 2x250bp reads. To confirm the transfer of pFIJ0137_1, we individually recruited reads from the *B. fragilis* 214 donor, naive *B. fragilis* 638R, and *B. fragilis* 638R transconjugants to the pFIJ0137_1 reference sequence. We used anvi'o to create contigs and profile databases (as described above) and visualized these results with the command `anvi-interactive`. We independently confirmed the presence of pFIJ0137_1 by assembling genomes using SPAdes[141] with default parameters.

Data availability

Reproducible analyses of reference plasmids and chromosomes are available at doi:10.5281/zenodo.5732024. The PlasX model as well as our analyses of known and predicted plasmids are available at doi:10.5281/zenodo.5843600. For all metagenomes, we have compiled the contigs, taxonomic abundances, and PlasX scores at doi:10.5281/zenodo.8175278, gene calls at doi:10.5281/zenodo.5730987, and gene annotations at doi:10.5281/zenodo.5731658. We have deposited long and short sequencing reads from *B. fragilis* isolates into the NCBI Sequence Read Archive (PRJNA782184).

Code availability

We have released two open-source packages, PlasX (https://github.com/michaelkyu/plasx) and MobMess (https://github.com/michaelkyu/mobmess), along with detailed installation and usage instructions.

Acknowledgments

We thank Karen Lolans (ORCiD:0000-0003-1903-756X) for performing the long-read sequencing and for providing feedback on the manuscript. We thank Samuel Miller (0000-0002-2836-1401) and Marcus Foo (0000-0003-3436-1632) for their insights into tRNA modification genes. We also thank other members of the Meren Lab at the University of Chicago for their feedback. MKY acknowledges support from Toyota Technological Institute at Chicago.

Funding

Center for Data and Computing, at the University of Chicago (MKY, AME)
National Institutes of Health NIDDK grant RC2 DK122394 (AME)
Simons Foundation grant #687269 (AME)
Sloan Foundation (AME)

Author information

Contributions

Conceptualization: MKY, AME
Methodology: MKY, ECF, AME
Investigation: MKY, ECF, AME
Visualization: MKY, ECF, AME
Funding acquisition: MKY, AME
Project administration: MKY, AME
Supervision: MKY, AME
Writing – original draft: MKY, ECF
Writing – review & editing: MKY, ECF, AME
Data curation: MKY, ECF
Formal Analysis: MKY, ECF
Resources: ECF, MKY
Software: MKY
Validation: ECF

Ethics declarations

Competing interests

The authors declare no competing interests.

References

1. Frost, L. S., Leplae, R., Summers, A. O. & Toussaint, A. Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.* **3,** 722–732 (2005).
2. Kumar, D., Prajapati, H. K., Mahilkar, A., Ma, C.-H., Mittal, P., Jayaram, M. & Ghosh, S. K. The selfish yeast plasmid utilizes the condensin complex and condensed chromatin for faithful partitioning. *PLoS Genet.* **17,** e1009660 (2021).
3. Kazlauskas, D., Varsani, A., Koonin, E. V. & Krupovic, M. Multiple origins of prokaryotic and eukaryotic single-stranded DNA viruses from bacterial and archaeal plasmids. *Nat. Commun.* **10,** 1–12 (2019).
4. del Solar, G., Giraldo, R., Ruiz-Echevarría, M. J., Espinosa, M. & Díaz-Orejas, R. Replication and Control of Circular Bacterial Plasmids. *Microbiol. Mol. Biol. Rev.* **62,** 434 (1998).
5. Khan, S. A. Rolling-circle replication of bacterial plasmids. *Microbiology and molecular biology reviews : MMBR* **61,** 442–455 Preprint at https://doi.org/10.1128/.61.4.442-455.1997 (1997)
6. Lilly, J. & Camps, M. Mechanisms of Theta Plasmid Replication. *Microbiol Spectr* **3,** PLAS–0029–2014 (2015).
7. Thomas, C. M. *Horizontal Gene Pool: Bacterial Plasmids and Gene Spread*. (CRC Press, 2003).
8. Summers, D. *The Biology of Plasmids*. (1993).
9. Jacob, A. E. & Hobbs, S. J. Conjugal transfer of plasmid-borne multiple antibiotic resistance in Streptococcus faecalis var. zymogenes. *J. Bacteriol.* **117,** 360–372 (1974).
10. Poyart-Salmeron, C., Carlier, C., Trieu-Cuot, P., Courtieu, A. L. & Courvalin, P. Transferable plasmid-mediated antibiotic resistance in Listeria monocytogenes. *Lancet* **335,** 1422–1426 (1990).

11. Lan, R., Stevenson, G. & Reeves, P. R. Comparison of Two Major Forms of the Shigella Virulence Plasmid pINV: Positive Selection Is a Major Force Driving the Divergence. *Infect. Immun.* **71,** 6298 (2003).

12. Meletzus, D., Bermphol, A., Dreier, J. & Eichenlaub, R. Evidence for plasmid-encoded virulence factors in the phytopathogenic bacterium Clavibacter michiganensis subsp. michiganensis NCPPB382. *J. Bacteriol.* **175,** 2131–2136 (1993).

13. Sen, D., Van der Auwera, G. A., Rogers, L. M., Thomas, C. M., Brown, C. J. & Top, E. M. Broad-host-range plasmids from agricultural soils have IncP-1 backbones with diverse accessory genes. *Appl. Environ. Microbiol.* **77,** 7975–7983 (2011).

14. Holt, K. E., Thomson, N. R., Wain, J., Phan, M. D., Nair, S., Hasan, R., Bhutta, Z. A., Quail, M. A., Norbertczak, H., Walker, D., Dougan, G. & Parkhill, J. Multidrug-resistant Salmonella enterica serovar paratyphi A harbors IncHI1 plasmids similar to those found in serovar typhi. *J. Bacteriol.* **189,** 4257–4264 (2007).

15. Fernandez-Lopez, R., Redondo, S., Garcillan-Barcia, M. P. & de la Cruz, F. Towards a taxonomy of conjugative plasmids. *Curr. Opin. Microbiol.* **38,** 106–113 (2017).

16. Oliva, M., Calia, C., Ferrara, M., D'Addabbo, P., Scrascia, M., Mulè, G., Monno, R. & Pazzani, C. Antimicrobial resistance gene shuffling and a three-element mobilisation system in the monophasic Salmonella typhimurium strain ST1030. *Plasmid* **111,** 102532 Preprint at https://doi.org/10.1016/j.plasmid.2020.102532 (2020)

17. Orlek, A., Stoesser, N., Anjum, M. F., Doumith, M., Ellington, M. J., Peto, T., Crook, D., Woodford, N., Walker, A. S., Phan, H. & Sheppard, A. E. Plasmid Classification in an Era of Whole-Genome Sequencing: Application in Studies of Antibiotic Resistance Epidemiology. *Front. Microbiol.* **0,** (2017).

18. Norberg, P., Bergström, M., Jethava, V., Dubhashi, D. & Hermansson, M. The IncP-1 plasmid backbone adapts to different host bacterial species and evolves through homologous recombination. *Nat. Commun.* **2,** 268 (2011).

19. Heuer, H. & Smalla, K. Plasmids foster diversification and adaptation of bacterial populations in soil. *FEMS Microbiol. Rev.* **36,** 1083–1104 (2012).
20. Sota, M., Yano, H., Hughes, J. M., Daughdrill, G. W., Abdo, Z., Forney, L. J. & Top, E. M. Shifts in the host range of a promiscuous plasmid through parallel evolution of its replication initiation protein. *ISME J.* **4,** 1568–1580 (2010).
21. Handelsman, J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* **68,** 669–685 (2004).
22. Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M. & Banfield, J. F. Accurate and complete genomes from metagenomes. *Genome Res.* **30,** 315–333 (2020).
23. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486,** 207–214 (2012).
24. Manor, O., Dai, C. L., Kornilov, S. A., Smith, B., Price, N. D., Lovejoy, J. C., Gibbons, S. M. & Magis, A. T. Health and disease markers correlate with gut microbiome composition across thousands of people. *Nat. Commun.* **11,** 1–12 (2020).
25. de Vos, W. M., Tilg, H., Van Hul, M. & Cani, P. D. Gut microbiome and health: mechanistic insights. *Gut* **71,** 1020–1032 (2022).
26. Smalla, K., Jechalke, S. & Top, E. M. Plasmid detection, characterization and ecology. *Microbiology spectrum* **3,** (2015).
27. Jones, B. V. & Marchesi, J. R. Transposon-aided capture (TRACA) of plasmids resident in the human gut mobile metagenome. *Nat. Methods* **4,** 55–61 (2007).
28. Delaney, S., Murphy, R. & Walsh, F. A Comparison of Methods for the Extraction of Plasmids Capable of Conferring Antibiotic Resistance in a Human Pathogen From Complex Broiler Cecal Samples. *Front. Microbiol.* **9,** 1731 (2018).
29. Brown Kav, A., Sasson, G., Jami, E., Doron-Faigenboim, A., Benhar, I. & Mizrahi, I. Insights into the bovine rumen plasmidome. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 5452–5457 (2012).
30. Krawczyk, P. S., Lipinski, L. & Dziembowski, A. PlasFlow: predicting plasmid sequences in

metagenomic data using genome signatures. *Nucleic Acids Res.* **46,** e35 (2018).

31. Pellow, D., Mizrahi, I. & Shamir, R. PlasClass improves plasmid sequence classification. *PLoS Comput. Biol.* **16,** e1007781 (2020).
32. Antipov, D., Raiko, M., Lapidus, A. & Pevzner, P. A. Plasmid detection and assembly in genomic and metagenomic data sets. *Genome Res.* **29,** 961–968 (2019).
33. Arredondo-Alonso, S., Willems, R. J., van Schaik, W. & Schürch, A. C. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microb Genom* **3,** e000128 (2017).
34. Andreopoulos, W. B., Geller, A. M., Lucke, M., Balewski, J., Clum, A., Ivanova, N. N. & Levy, A. Deeplasmid: deep learning accurately separates plasmids from bacterial chromosomes. *Nucleic Acids Res.* **50,** e17 (2022).
35. Hou, S., Cheng, S., Chen, T., Fuhrman, J. A. & Sun, F. DeepMicrobeFinder Sorts Metagenomes into Prokaryotes, Eukaryotes and Viruses, with Marine Applications. *Research Square* (2021). doi:10.21203/rs.3.rs-1016976/v1
36. Zhou, F. & Xu, Y. cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics* **26,** 2051–2052 (2010).
37. Carattoli, A., Zankari, E., García-Fernández, A., Larsen, M. V., Lund, O., Villa, L., Aarestrup, F. M. & Hasman, H. In SilicoDetection and Typing of Plasmids using PlasmidFinder and Plasmid Multilocus Sequence Typing. *Antimicrobial Agents and Chemotherapy* **58,** 3895–3903 Preprint at https://doi.org/10.1128/aac.02412-14 (2014)
38. Robertson, J. & Nash, J. H. E. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom* **4,** (2018).
39. Garcillán-Barcia, M. P., Francia, M. V. & de la Cruz, F. The diversity of conjugative relaxases and its application in plasmid classification. *FEMS Microbiol. Rev.* **33,** 657–687 (2009).

40. Rozov, R., Brown Kav, A., Bogumil, D., Shterzer, N., Halperin, E., Mizrahi, I. & Shamir, R. Recycler: an algorithm for detecting plasmids from de novo assembly graphs. *Bioinformatics* **33,** 475–482 (2017).
41. Pellow, D., Zorea, A., Probst, M., Furman, O., Segal, A., Mizrahi, I. & Shamir, R. SCAPP: an algorithm for improved plasmid assembly in metagenomes. *Microbiome* **9,** 144 (2021).
42. Galperin, M. Y., Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.* **43,** D261–9 (2015).
43. El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E. & Finn, R. D. The Pfam protein families database in 2019. *Nucleic Acids Res.* **47,** D427–D432 (2019).
44. Fang, Z., Tan, J., Wu, S., Li, M., Xu, C., Xie, Z. & Zhu, H. PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. *Gigascience* **8,** (2019).
45. Schwengers, O., Barth, P., Falgenhauer, L., Hain, T., Chakraborty, T. & Goesmann, A. Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores. *Microb Genom* **6,** (2020).
46. Schmartz, G. P., Hartung, A., Hirsch, P., Kern, F., Fehlmann, T., Müller, R. & Keller, A. PLSDB: advancing a comprehensive database of bacterial plasmids. *Nucleic Acids Res.* **50,** D273–D278 (2022).
47. Reveillaud, J., Bordenstein, S. R., Cruaud, C., Shaiber, A., Esen, Ö. C., Weill, M., Makoundou, P., Lolans, K., Watson, A. R., Rakotoarivony, I., Bordenstein, S. R. & Eren, A. M. The Wolbachia mobilome in Culex pipiens includes a putative plasmid. *Nat. Commun.* **10,** 1–11 (2019).

1235 48. Liu, M., Li, X., Xie, Y., Bi, D., Sun, J., Li, J., Tai, C., Deng, Z. & Ou, H.-Y. ICEberg 2.0: an updated database of bacterial integrative and conjugative elements. *Nucleic Acids Res.* **47,** D660–D665 (2019).

49. Jalal, A. S. B. & Le, T. B. K. Bacterial chromosome segregation by the ParABS system. *Open Biol.* **10,** 200097 (2020).

1240 50. Bouet, J.-Y. & Funnell, B. E. Plasmid Localization and Partition in Enterobacteriaceae. *EcoSal Plus* **8,** (2019).

51. Sukupolvi, S. & O'Connor, C. D. TraT lipoprotein, a plasmid-specified mediator of interactions between gram-negative bacteria and their environment. *Microbiol. Rev.* **54,** 331–341 (1990).

1245 52. Norris, S. J., Carter, C. J., Howell, J. K. & Barbour, A. G. Low-passage-associated proteins of Borrelia burgdorferi B31: characterization and molecular cloning of OspD, a surface-exposed, plasmid-encoded lipoprotein. *Infect. Immun.* **60,** 4662–4672 (1992).

53. Carr, V. R., Shkoporov, A., Hill, C., Mullany, P. & Moyes, D. L. Probing the Mobilome: Discoveries in the Dynamic Microbiome. *Trends Microbiol.* **29,** 158–170 (2021).

1250 54. Meinhardt, F., Schaffrath, R. & Larsen, M. Microbial linear plasmids. *Appl. Microbiol. Biotechnol.* **47,** 329–336 (1997).

55. Vineis, J. H., Ringus, D. L., Morrison, H. G., Delmont, T. O., Dalal, S., Raffals, L. H., Antonopoulos, D. A., Rubin, D. T., Eren, A. M., Chang, E. B. & Sogin, M. L. Patient-Specific Bacteroides Genome Variants in Pouchitis. *MBio* **7,** (2016).

1255 56. Monaghan, T. M., Sloan, T. J., Stockdale, S. R., Blanchard, A. M., Emes, R. D., Wilcox, M., Biswas, R., Nashine, R., Manke, S., Gandhi, J., Jain, P., Bhotmange, S., Ambalkar, S., Satav, A., Draper, L. A., Hill, C. & Kashyap, R. S. Metagenomics reveals impact of geography and acute diarrheal disease on the Central Indian human gut microbiome. *Gut Microbes* **12,** 1752605 (2020).

1260 57. Mancabelli, L., Milani, C., Lugli, G. A., Turroni, F., Ferrario, C., van Sinderen, D. & Ventura,

M. Meta-analysis of the human gut microbiome from urbanized and pre-agricultural populations. *Environ. Microbiol.* **19,** 1379–1390 (2017).

58. Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., Collado, M. C., Rice, B. L., DuLong, C., Morgan, X. C., Golden, C. D., Quince, C., Huttenhower, C. & Segata, N. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176,** 649–662.e20 (2019).
59. Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R. N., Anokhin, A. P., Heath, A. C., Warner, B., Reeder, J., Kuczynski, J., Caporaso, J. G., Lozupone, C. A., Lauber, C., Clemente, J. C., Knights, D., Knight, R. & Gordon, J. I. Human gut microbiome viewed across age and geography. *Nature* **486,** 222–227 (2012).
60. Klümper, U., Riber, L., Dechesne, A., Sannazzarro, A., Hansen, L. H., Sørensen, S. J. & Smets, B. F. Broad host range plasmids can invade an unexpectedly diverse fraction of a soil bacterial community. *ISME J.* **9,** (2015).
61. Kohler, V., Vaishampayan, A. & Grohmann, E. Broad-host-range Inc18 plasmids: Occurrence, spread and transfer mechanisms. *Plasmid* **99,** 11–21 (2018).
62. Bishé, B., Taton, A. & Golden, J. W. Modification of RSF1010-Based Broad-Host-Range Plasmids for Improved Conjugation and Cyanobacterial Bioprospecting. *iScience* **20,** 216–228 (2019).
63. Lloyd-Price, J., Arze, C., Ananthakrishnan, A. N., Schirmer, M., Avila-Pacheco, J., Poon, T. W., Andrews, E., Ajami, N. J., Bonham, K. S., Brislawn, C. J., Casero, D., Courtney, H., Gonzalez, A., Graeber, T. G., Hall, A. B., Lake, K., Landers, C. J., Mallick, H., Plichta, D. R., Prasad, M., Rahnavard, G., Sauk, J., Shungin, D., Vázquez-Baeza, Y., White, R. A., 3rd, IBDMDB Investigators, Braun, J., Denson, L. A., Jansson, J. K., Knight, R., Kugathasan, S., McGovern, D. P. B., Petrosino, J. F., Stappenbeck, T. S., Winter, H. S.,

Clish, C. B., Franzosa, E. A., Vlamakis, H., Xavier, R. J. & Huttenhower, C. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569,** 655–662 (2019).

64. Schloss, P. D. Identifying and Overcoming Threats to Reproducibility, Replicability, Robustness, and Generalizability in Microbiome Research. *MBio* **9,** (2018).
65. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018). at <http://arxiv.org/abs/1802.03426>
66. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8,** e1002687 (2012).
67. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **3,** e104 (2017).
68. Fernández-López, R., Garcillán-Barcia, M. P., Revilla, C., Lázaro, M., Vielva, L. & de la Cruz, F. Dynamics of the IncW genetic backbone imply general trends in conjugative plasmid evolution. *FEMS Microbiol. Rev.* **30,** 942–966 (2006).
69. Garcillán-Barcia, M. P., Ruiz del Castillo, B., Alvarado, A., de la Cruz, F. & Martínez-Martínez, L. Degenerate primer MOB typing of multiresistant clinical isolates of E. coli uncovers new plasmid backbones. *Plasmid* **77,** 17–27 (2015).
70. Carattoli, A., Bertini, A., Villa, L., Falbo, V., Hopkins, K. L. & John Threlfall, E. Identification of plasmids by PCR-based replicon typing. *Journal of Microbiological Methods* **63,** 219–228 Preprint at https://doi.org/10.1016/j.mimet.2005.03.018 (2005)
71. Carattoli, A., Zankari, E., García-Fernández, A., Voldby Larsen, M., Lund, O., Villa, L., Møller Aarestrup, F. & Hasman, H. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.* **58,** 3895–3903 (2014).
72. Redondo-Salvo, S., Fernández-López, R., Ruiz, R., Vielva, L., de Toro, M., Rocha, E. P. C., Garcillán-Barcia, M. P. & de la Cruz, F. Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. *Nat. Commun.* **11,** 3602 (2020).

73. Acman, M., van Dorp, L., Santini, J. M. & Balloux, F. Large-scale network analysis captures biological features of bacterial plasmids. *Nat. Commun.* **11,** 1–11 (2020).

74. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–2504 (2003).

75. Heer, J., Card, S. K. & Landay, J. A. prefuse: a toolkit for interactive information visualization. in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 421–430 (Association for Computing Machinery, 2005).

76. Vrancianu, C. O., Popa, L. I., Bleotu, C. & Chifiriuc, M. C. Targeting Plasmids to Limit Acquisition and Transmission of Antimicrobial Resistance. *Front. Microbiol.* **11,** 761 (2020).

77. MacLean, R. C. & San Millan, A. The evolution of antibiotic resistance. *Science* **365,** 1082–1083 (2019).

78. World Health Organization. *WHO report on surveillance of antibiotic consumption: 2016-2018 early implementation*. (World Health Organization, 2018). at <https://apps.who.int/iris/bitstream/handle/10665/277359/9789241514880-eng.pdf>

79. Centers for Disease Control and Prevention. *Antibiotic Use in the United States, 2021 Update: Progress and Opportunities*. (2021). at <https://www.cdc.gov/antibiotic-use/stewardship-report/current.html>

80. Gehrig, S., Eberle, M.-E., Botschen, F., Rimbach, K., Eberle, F., Eigenbrod, T., Kaiser, S., Holmes, W. M., Erdmann, V. A., Sprinzl, M., Bec, G., Keith, G., Dalpke, A. H. & Helm, M. Identification of modifications in microbial, native tRNA that suppress immunostimulatory activity. *J. Exp. Med.* **209,** 225–233 (2012).

81. Galvanin, A., Vogt, L.-M., Grober, A., Freund, I., Ayadi, L., Bourguignon-Igel, V., Bessler, L., Jacob, D., Eigenbrod, T., Marchand, V., Dalpke, A., Helm, M. & Motorin, Y. Bacterial tRNA 2'-O-methylation is dynamically regulated under stress conditions and modulates innate immune response. *Nucleic Acids Res.* **48,** 12833–12844 (2020).

82. Embers, M. E., Alvarez, X., Ooms, T. & Philipp, M. T. The failure of immune response
1340 evasion by linear plasmid 28-1-deficient Borrelia burgdorferi is attributable to persistent expression of an outer surface protein. *Infect. Immun.* **76,** 3984–3991 (2008).

83. Svara, F. & Rankin, D. J. The evolution of plasmid-carried antibiotic resistance. *BMC Evol. Biol.* **11,** 130 (2011).

84. Sykes, R. The 2009 Garrod lecture: the evolution of antimicrobial resistance: a Darwinian
1345 perspective. *J. Antimicrob. Chemother.* **65,** 1842–1852 (2010).

85. Cantón, R. & Morosini, M.-I. Emergence and spread of antibiotic resistance following exposure to antibiotics. *FEMS Microbiol. Rev.* **35,** 977–991 (2011).

86. Baquero, F. Low-level antibacterial resistance: a gateway to clinical resistance. *Drug Resist. Updat.* **4,** 93–105 (2001).

1350 87. San Millan, A., Escudero, J. A., Gifford, D. R., Mazel, D. & MacLean, R. C. Multicopy plasmids potentiate the evolution of antibiotic resistance in bacteria. *Nat Ecol Evol* **1,** 10 (2016).

88. Alonso, A., Sanchez, P. & Martinez, J. L. Environmental selection of antibiotic resistance genes. Minireview. *Environmental Microbiology* **3,** 1–9 Preprint at
1355 https://doi.org/10.1046/j.1462-2920.2001.00161.x (2001)

89. Xiong, W., Sun, Y., Ding, X., Wang, M. & Zeng, Z. Selective pressure of antibiotics on ARGs and bacterial communities in manure-polluted freshwater-sediment microcosms. *Front. Microbiol.* **0,** (2015).

90. Ma, H. & Bryers, J. D. Non-invasive determination of conjugative transfer of plasmids
1360 bearing antibiotic-resistance genes in biofilm-bound bacteria: effects of substrate loading and antibiotic selection. *Appl. Microbiol. Biotechnol.* **97,** 317–328 (2012).

91. Berendsen, B., Stolker, L., de Jong, J., Nielen, M., Tserendorj, E., Sodnomdarjaa, R., Cannavan, A. & Elliott, C. Evidence of natural occurrence of the banned antibiotic chloramphenicol in herbs and grass. *Anal. Bioanal. Chem.* **397,** 1955 (2010).

1365 92. Both, L., Botgros, R. & Cavaleri, M. Analysis of licensed over-the-counter (OTC) antibiotics in the European Union and Norway, 2012. *Euro Surveill.* **20,** 30002 (2015).

93. Balbi, H. J. Chloramphenicol: A Review. *Pediatrics in Review* **25,** 284–288 Preprint at https://doi.org/10.1542/pir.25-8-284 (2004)

94. Ministry of Health and Medical Services, Government of Fiji. *Fiji Antibiotic Guidelines*.
1370 (Government of Fiji, 2019).

95. Royer, G., Decousser, J. W., Branger, C., Dubois, M., Médigue, C., Denamur, E. & Vallenet, D. PlaScope: a targeted approach to assess the plasmidome from genome assemblies at the species level. *Microb Genom* **4,** (2018).

96. Arredondo-Alonso, S., Rogers, M. R. C., Braat, J. C., Verschuuren, T. D., Top, J.,
1375 Corander, J., Willems, R. J. L. & Schürch, A. C. Mlplasmids: A user-friendly tool to predict plasmid- and chromosome-derived sequences for single species. *Microb. Genom.* **4,** (2018).

97. Gomi, R., Wyres, K. L. & Holt, K. E. Detection of plasmid contigs in draft genome assemblies using customized Kraken databases. *Microb Genom* **7,** (2021).

1380 98. Wozniak, R. A. F. & Waldor, M. K. Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. *Nat. Rev. Microbiol.* **8,** 552–563 (2010).

99. Wang, G. H., Sun, B. F., Xiong, T. L., Wang, Y. K., Murfin, K. E., Xiao, J. H. & Huang, D. W. Bacteriophage WO Can Mediate Horizontal Gene Transfer in Endosymbiotic Wolbachia
1385 Genomes. *Front. Microbiol.* **0,** (2016).

100. Cuecas, A., Kanoksilapatham, W. & Gonzalez, J. M. Evidence of horizontal gene transfer by transposase gene analyses in Fervidobacterium species. *PLoS One* **12,** e0173961 (2017).

101. Garcillán-Barcia, M. P., Alvarado, A. & de la Cruz, F. Identification of bacterial plasmids
1390 based on mobility and plasmid population biology. *FEMS Microbiol. Rev.* **35,** 936–956

(2011).

102. Galata, V., Fehlmann, T., Backes, C. & Keller, A. PLSDB: a resource of complete bacterial plasmids. *Nucleic Acids Res.* **47,** D195–D202 (2019).

103. Shaiber, A., Willis, A. D., Delmont, T. O., Roux, S., Chen, L.-X., Schmid, A. C., Yousef, M., Watson, A. R., Lolans, K., Esen, Ö. C., Lee, S. T. M., Downey, N., Morrison, H. G., Dewhirst, F. E., Mark Welch, J. L. & Eren, A. M. Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. *Genome Biol.* **21,** 292 (2020).

104. Eren, A. M., Kiefl, E., Shaiber, A., Veseli, I., Miller, S. E., Schechter, M. S., Fink, I., Pan, J. N., Yousef, M., Fogarty, E. C., Trigodet, F., Watson, A. R., Esen, Ö. C., Moore, R. M., Clayssen, Q., Lee, M. D., Kivenson, V., Graham, E. D., Merrill, B. D., Karkman, A., Blankenberg, D., Eppley, J. M., Sjödin, A., Scott, J. J., Vázquez-Campos, X., McKay, L. J., McDaniel, E. A., Stevens, S. L. R., Anderson, R. E., Fuessel, J., Fernandez-Guerra, A., Maignien, L., Delmont, T. O. & Willis, A. D. Community-led, integrated, reproducible multi-omics with anvi'o. *Nat Microbiol* **6,** 3–6 (2021).

105. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28,** 2520–2522 (2012).

106. Hyatt, D., Chen, G.-L., Locascio, P. F., Land, M. L., Larimer, F. W. & Hauser, L. J. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11,** 119 (2010).

107. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12,** 59–60 (2015).

108. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7,** e1002195 (2011).

109. Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nat. Commun.* **9,** 1–8 (2018).

110. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high

throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).

111. Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A. & Lanfear, R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37,** 1530–1534 (2020).

112. Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S. & Phillippy, A. M. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17,** 132 (2016).

113. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. & Others. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12,** 2825–2830 (2011).

114. Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., Zhang, D., Xia, H., Xu, X., Jie, Z., Su, L., Li, X., Li, X., Li, J., Xiao, L., Huber-Schönauer, U., Niederseer, D., Xu, X., Al-Aama, J. Y., Yang, H., Wang, J., Kristiansen, K., Arumugam, M., Tilg, H., Datz, C. & Wang, J. Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nat. Commun.* **6,** 1–13 (2015).

115. David, L. A., Weil, A., Ryan, E. T., Calderwood, S. B., Harris, J. B., Chowdhury, F., Begum, Y., Qadri, F., LaRocque, R. C. & Turnbaugh, P. J. Gut Microbial Succession Follows Acute Secretory Diarrhea in Humans. *MBio* **6,** (2015).

116. Raymond, F., Ouameur, A. A., Déraspe, M., Iqbal, N., Gingras, H., Dridi, B., Leprohon, P., Plante, P.-L., Giroux, R., Bérubé, È., Frenette, J., Boudreau, D. K., Simard, J.-L., Chabot, I., Domingo, M.-C., Trottier, S., Boissinot, M., Huletsky, A., Roy, P. H., Ouellette, M., Bergeron, M. G. & Corbeil, J. The initial state of the human gut microbiome determines its reshaping by antibiotics. *ISME J.* **10,** 707 (2016).

117. Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.-M., Hansen, T., Le

Paslier, D., Linneberg, A., Nielsen, H. B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Doré, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Bork, P., Ehrlich, S. D. & Wang, J. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464,** 59–65 (2010).

118. Wen, C., Zheng, Z., Shao, T., Liu, L., Xie, Z., Le Chatelier, E., He, Z., Zhong, W., Fan, Y., Zhang, L., Li, H., Wu, C., Hu, C., Xu, Q., Zhou, J., Cai, S., Wang, D., Huang, Y., Breban, M., Qin, N. & Ehrlich, S. D. Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis. *Genome Biol.* **18,** (2017).

119. Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., Almeida, M., Arumugam, M., Batto, J. M., Kennedy, S., Leonard, P., Li, J., Burgdorf, K., Grarup, N., Jørgensen, T., Brandslund, I., Nielsen, H. B., Juncker, A. S., Bertalan, M., Levenez, F., Pons, N., Rasmussen, S., Sunagawa, S., Tap, J., Tims, S., Zoetendal, E. G., Brunak, S., Clément, K., Doré, J., Kleerebezem, M., Kristiansen, K., Renault, P., Sicheritz-Ponten, T., de Vos, W. M., Zucker, J. D., Raes, J., Hansen, T., Bork, P., Wang, J., Ehrlich, S. D. & Pedersen, O. Richness of human gut microbiome correlates with metabolic markers. *Nature* **500,** (2013).

120. Xie, H., Guo, R., Zhong, H., Feng, Q., Lan, Z., Qin, B., Ward, K. J., Jackson, M. A., Xia, Y., Chen, X., Chen, B., Xia, H., Xu, C., Li, F., Xu, X., Al-Aama, J. Y., Yang, H., Wang, J., Kristiansen, K., Wang, J., Steves, C. J., Bell, J. T., Li, J., Spector, T. D. & Jia, H. Shotgun Metagenomics of 250 Adult Twins Reveals Genetic and Environmental Impacts on the Gut Microbiome. *Cell systems* **3,** 572 (2016).

121. Brito, I. L., Yilmaz, S., Huang, K., Xu, L., Jupiter, S. D., Jenkins, A. P., Naisilisili, W., Tamminen, M., Smillie, C. S., Wortman, J. R., Birren, B. W., Xavier, R. J., Blainey, P. C., Singh, A. K., Gevers, D. & Alm, E. J. Mobile genes in the human microbiome are structured from global to individual scales. *Nature* **535,** 435 (2016).

122. Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., Ben-Yacov, O., Lador, D., Avnit-Sagi, T., Lotan-Pompan, M., Suez, J., Mahdi, J. A., Matot, E., Malka, G., Kosower, N., Rein, M., Zilberman-Schapira, G., Dohnalová, L., Pevsner-Fischer, M., Bikovsky, R., Halpern, Z., Elinav, E. & Segal, E. Personalized Nutrition by Prediction of Glycemic Responses. *Cell* **163,** 1079–1094 (2015).

123. Rampelli, S., Schnorr, S. L., Consolandi, C., Turroni, S., Severgnini, M., Peano, C., Brigidi, P., Crittenden, A. N., Henry, A. G. & Candela, M. Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota. *Curr. Biol.* **25,** 1682–1693 (2015).

124. Liu, W., Zhang, J., Wu, C., Cai, S., Huang, W., Chen, J., Xiaoxia, X. I., Liang, Z., Hou, Q., Zhou, B., Qin, N. & Zhang, H. Unique Features of Ethnic Mongolian Gut Microbiome revealed by metagenomic analysis. *Sci. Rep.* **6,** (2016).

125. Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., Arumugam, M., Kultima, J. R., Prifti, E., Nielsen, T., Juncker, A. S., Manichanh, C., Chen, B., Zhang, W., Levenez, F., Wang, J., Xu, X., Xiao, L., Liang, S., Zhang, D., Zhang, Z., Chen, W., Zhao, H., Al-Aama, J. Y., Edris, S., Yang, H., Wang, J., Hansen, T., Nielsen, H. B., Brunak, S., Kristiansen, K., Guarner, F., Pedersen, O., Doré, J., Ehrlich, S. D., Bork, P. & Wang, J. An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32,** 834–841 (2014).

126. Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R. & Gordon, J. I. The Human Microbiome Project. *Nature* **449,** 804–810 (2007).

127. Obregon-Tito, A. J., Tito, R. Y., Metcalf, J., Sankaranarayanan, K., Clemente, J. C., Ursell, L. K., Xu, Z. Z., Van Treuren, W., Knight, R., Gaffney, P. M., Spicer, P., Lawson, P., Marin-Reyes, L., Trujillo-Villarroel, O., Foster, M., Guija-Poma, E., Troncoso-Corzo, L., Warinner, C., Ozga, A. T. & Lewis, C. M. Subsistence strategies in traditional societies distinguish gut microbiomes. *Nat. Commun.* **6,** 1–9 (2015).

128. Shaiber, A. & Murat Eren, A. Anvi'o snakemake workflows. (2018). at

1495 <http://merenlab.org/2018/07/09/anvio-snakemake-workflows/>

129. Murat Eren, A., Vineis, J. H., Morrison, H. G. & Sogin, M. L. A Filtering Method to Generate High Quality Short Reads Using Illumina Paired-End Technology. *PLoS One* **8,** e66643 (2013).

130. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for
1500 single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28,** 1420–1428 (2012).

131. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357 (2012).

132. Jørgensen, T. S., Xu, Z., Hansen, M. A., Sørensen, S. J. & Hansen, L. H. Hundreds of
1505 circular novel plasmids and DNA elements identified in a rat cecum metamobilome. *PLoS One* **9,** e87924 (2014).

133. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20,** 1–13 (2019).

134. Watts, S. C., Ritchie, S. C., Inouye, M. & Holt, K. E. FastSpar: rapid and scalable
1510 correlation estimation for compositional data. *Bioinformatics* **35,** 1064–1066 (2018).

135. Utter, D. R., Borisy, G. G., Eren, A. M., Cavanaugh, C. M. & Mark Welch, J. L. Metapangenomics of the oral microbiome provides insights into habitat adaptation and cultivar diversity. *Genome Biol.* **21,** 293 (2020).

136. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High
1515 throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9,** 5114 (2018).

137. Csardi, G., Nepusz, T. & Others. The igraph software package for complex network research. *InterJournal, complex systems* **1695,** 1–9 (2006).

138. Gibson, M. K., Forsberg, K. J. & Dantas, G. Improved annotation of antibiotic resistance
1520 determinants reveals microbial resistomes cluster by ecology. *ISME J.* **9,** 207–216 (2015).

139. Alcock, B. P., Raphenya, A. R., Lau, T. T. Y., Tsang, K. K., Bouchard, M., Edalatmand, A., Huynh, W., Nguyen, A.-L. V., Cheng, A. A., Liu, S., Min, S. Y., Miroshnichenko, A., Tran, H.-K., Werfalli, R. E., Nasir, J. A., Oloni, M., Speicher, D. J., Florescu, A., Singh, B., Faltyn, M., Hernandez-Koutoucheva, A., Sharma, A. N., Bordeleau, E., Pawlowski, A. C., Zubyk, H. L., Dooley, D., Griffiths, E., Maguire, F., Winsor, G. L., Beiko, R. G., Brinkman, F. S. L., Hsiao, W. W. L., Domselaar, G. V. & McArthur, A. G. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **48,** D517–D525 (2020).

140. Trigodet, F., Lolans, K. & Fogarty, E. High molecular weight DNA extraction strategies for long-read sequencing of complex metagenomes. *Mol. Ecol.* (2022). at <https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13588>

141. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A. & Pevzner, P. A. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19,** 455 (2012).